Pairs Trading Strategy Design and Backtest (TS)

Julien Granger, Ph.D.[1]

[1]juliengranger@gmail.com

January 23, 2023

Abstract

We design a pairs trading strategy for oil futures. We implement regression estimation in matrix form, a VAR function, a Granger causality test, the Engle-Granger procedure from first principles, the Johansen procedure from off-the-shelf libraries, the Kalman filter from first principles, and a backtest engine (in two flavors: whole horizon and rolling estimation-window/out-of-sample backtest) in the python programming language. This allows us to study calendar spreads, intercommodity spreads and locational spreads among a universe of 37 pairs indexed over 1989-present and to identify those with the shortest half-lives. We compare spreads obtained via Engle-Granger, Johansen and Kalman approaches, report on trading results and preferred spreads.

Contents

1 Introduction

1.1 Industry background

Oil futures prices such as ICE Brent crude oil or NYMEX heating oil reflect the equilibrium of supply and demand balances in physical commodity markets, among other factors. Such commodity markets may experience temporary imbalanced situations such as a shortfall in demand (e.g. refinery turnarounds lead to reduced demand for crude oil) or a higher flow of supply (e.g. during the Covid-19 pandemic jet fuel was downgraded to the diesel fuel pool and thus put downward pressure on the heating oil futures contract).

However, such imbalances cannot persist for long durations because physical market participants will act upon them; therefore differentials in oil futures, such as calendar spreads and intercommodity spreads, tend to revert to their mean. That makes them good candidates for mean-reverting trading strategies based on statistical arbitrage (a.k.a. "spread trading" or "pairs trading"). In this project, we focus on pairs trading strategy design and backtest in energy futures.

1.2 Objectives

Our objectives in this study are to:

1. Provide a brief bibliographic review of spread trading for energy futures,
2. Provide a brief mathematical review of stationarity testing for time series, Vector Autoregression (VAR) modeling, time series cointegration and Vector Error Correction Model (VECM), and Kalman filtering,
3. Implement in python linear regression estimation in matrix form, the Engle-Granger procedure, the Johansen procedure (using off-the-self libraries), as well as the Kalman filter algorithm,
4. Design and backtest a mean-reverting trading strategy, using two approaches:
    4.1. First approach where all empirical data is used for estimation and backtesting,
    4.2. Second approach where a window of in-sample estimation/out-of-sample testing is rolled throughout the available history. This allows us to avoid hindsight bias for the performance evaluation of the Engle-Granger spread.

5. Identify cointegrated oil futures spreads, among timespreads, cracks and locational spreads and report on which provide the highest average annual return, Sharpe ratio and lowest maximum drawdown over our study period.

2 Data

Data used in this report include settlement prices for the futures of the nearest 12 months for the futures below, retrieved from Refinitiv Eikon, for the period 1989-present:

1. Brent Crude Oil (LCO),
2. WTI Crude Oil (CL),
3. Natural Gas (NG),
4. Heating Oil (HO),
5. Gasoline (RBOB, Reformulated gasoline Blendstock for Oxygenate Blending) (RB),
6. Gasoil (GO).

According to the data provider, the futures price series are constructed by rolling over on the first trading day of the month. This avoids the volatility typically observed near expiry; however it may create jumps in the series, for which we do not correct as these would be experienced in real trading practice of futures contracts. Except for Natural Gas, series are converted to be expressed in dollar per bbl. Following Alizadeh & Nomikos [1], price levels are transformed to natural logarithms. Note that each time we will be re-estimating cointegration afresh (at the start of each window), we normalize price levels to start at one, since cointegration analysis is known to be sensitive to initial conditions.

3 Bibliographic review

Spread trading among energy futures has been studied extensively in the literature.

Amizadeh & Nomikos [1] study cointegration of weekly price data for ICE Brent crude oil, NYMEX WTI crude oil, NYMEX Heating oil and ICE Gasoil using Vector Error Correction Modeling (VECM) and the Johansen methodology. They test moving average trading strategies and report Sharpe ratios as high as 2.1. They find that spreads including Gasoil in one leg generally tend to perform better (i.e. have the shortest half-lives, etc.). They also test the robustness of their strategy using the stationary bootstrap approach, which allows them to generate synthetic price data with statistically similar properties as the original empirical data.

Lubnau & Todorova [7] study cointegration using dynamic linear regression and Kalman filters for Brent, WTI, natural gas, RBOB (Reformulated gasoline Blendstock for Oxygenate Blending) and heating oil and moving average trading strategies. They find most combinations involving the front-month and second-month futures to be significantly profitable for all commodities tested, the best results for the Sharpe ratio are obtained for WTI and natural gas, with Sharpe ratios in excess of 2 for most combinations.

Cummins & Bocca [4] study spread trading among energy futures. They introduce a stepdown and balanced stepdown procedures to account for data snooping and minimize the risk of false discoveries.

Nakajima [8] studies cointegration between PJM Western hub wholesale electricity futures and Henry Hub natural gas futures using dynamic ordinary least squares solved via the Johansen procedure. The author uses a rolling three-year window to estimate the cointegration relationship. Interestingly, when the presence of cointegration is rejected, the author explains that a trader/hedger still may elect to trade, with the full knowledge that deviation from equilibrium may continue.

4 Mathematical review

Facts reviewed below are well-known and can be found in any graduate textbook on financial statistics such as Ruppert and Matteson [9], or Diamond [6].

4.1 Augmented Dickey-Fuller (ADF) unit root test

The ADF test is used for determining whether a single time series process has a unit root and thus is non-stationary. It is an extension of the basic Dickey-Fuller test as it includes higher-order auto-regressive terms in the test equation. Assuming that the series has a drift but no linear trend, the model equation is:

$$\Delta y_t = a_0 + \gamma y_{t-1} + \sum_{i=1}^{n} \beta_i \Delta y_{t-i} + \epsilon_t \tag{1}$$

where ϵ_t are i.i.d. with mean zero. Under the null hypothesis, $\gamma = 0$: the process has unit root, meaning the innovations have a permanent effect and the series is non-stationary. If the null hypothesis can be rejected in a one-sided test in favor of the alternative hypothesis $\gamma < 0$, the series is stationary and the innovations have a transitory effect. To reject the null, we use a test statistic $\tau_\mu = \frac{\hat{\gamma}}{SE(\hat{\gamma})}$ which has a specific distribution obtained by Dickey and Fuller as tabulated by McKinnon (2010 update).

4.2 Vector Autoregression (VAR)

Vector autoregression is the standard model to analyze a vector of stationary time series $\{\boldsymbol{Y_t}\}$. If the time series are not stationary, then we use a different type of analysis called cointegration, which we describe in subsequent sections.

A stationary VAR(p) is written as:

$$\boldsymbol{Y_t} = \boldsymbol{B_0} + \boldsymbol{B_1 Y_{t-1}} + \boldsymbol{B_2 Y_{t-2}} + ... + \boldsymbol{B_p Y_{t-p}} + \boldsymbol{\epsilon_t}, \tag{2}$$

where $\boldsymbol{\epsilon_t}$ is a vector white noise process.

By subtracting the mean and stacking p lags of $\boldsymbol{Y_t}$ into a large column vector $\boldsymbol{z_t}$, a VAR(p) is equivalently expressed as a VAR(1) using the companion form:

$$\boldsymbol{Z_t} = \boldsymbol{\Psi Z_{t-1}} + \boldsymbol{\chi_t}, \tag{3}$$

$$\boldsymbol{Z_t} = \begin{bmatrix} \boldsymbol{Y_t} - \mu \\ \boldsymbol{Y_{t-1}} - \mu \\ ... \\ \boldsymbol{Y_{t-p+1}} - \mu \end{bmatrix}, \tag{4}$$

$$\mu = \boldsymbol{B_0} \left(\boldsymbol{I} - \sum_{i=1}^{p} \boldsymbol{B_i} \right)^{-1}, \tag{5}$$

$$\boldsymbol{\Psi} = \begin{bmatrix} \boldsymbol{B_1} & \boldsymbol{B_2} & \boldsymbol{B_3} & ... & \boldsymbol{B_{p-1}} & \boldsymbol{B_p} \\ \boldsymbol{I_k} & \boldsymbol{0} & \boldsymbol{0} & ... & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I_k} & \boldsymbol{0} & ... & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & ... & \boldsymbol{I_k} & \boldsymbol{0} \end{bmatrix}, \tag{6}$$

$$\boldsymbol{\chi_t} = \begin{bmatrix} \boldsymbol{\epsilon_t} \\ \boldsymbol{0} \\ ... \\ \boldsymbol{0} \end{bmatrix}. \tag{7}$$

VAR(p) is covariance stationary (stable) if all of the eigenvalues of $\boldsymbol{\Psi}$ are less than one in absolute value (modulus if complex), which is a property we check in our python VAR implementation.

4.3 Granger causality

From our VAR implementation we are able to test for Granger causality via a likelihood ratio test. In the VAR(p)

$$\boldsymbol{Y_t} = \boldsymbol{B_0} + \boldsymbol{B_1 Y_{t-1}} + \boldsymbol{B_2 Y_{t-2}} + ... + \boldsymbol{B_p Y_{t-p}} + \boldsymbol{\epsilon_t}, \tag{8}$$

$Y_{j,t}$ does not Granger cause $Y_{i,t}$ if $B_{i,j,1} = B_{i,j,2} = ... = 0$ (restricted model). The likelihood ratio test statistic for testing the null $H_0 : B_{i,j,m} = 0, \forall m \in \{1, 2, .., p\}$ is as follows:

1. Estimate the unrestricted model $Y_i = x_i\beta + \epsilon_i$ and the restricted model $\bar{Y}_i = x_i\bar{\beta} + \bar{\epsilon}_i$,
2. Compute $SSE_R = \sum_{i=1}^{n} \bar{\epsilon_i}^2$ where $\bar{\epsilon_i}$ are the residuals from the restricted regression and $SSE_U = \sum_{i=1}^{n} \epsilon_i{}^2$ where ϵ_i are the residuals of the unrestricted regression,
3. Compute $LR = n \times \log\left(\frac{SSE_R}{SSE_U}\right)$,
4. Compute $W = \frac{n-k}{m}\left[\exp(\frac{LR}{n}) - 1\right]$ (in this case $m = 1$),
5. Compare W to the critical value C_α of the $F_{m,n-k}$ distribution at size α,
6. Reject the null (i.e. accept that Granger causality is present) if $W > C_\alpha$.

We implement this test in python using our VAR implementation and use it in our MyPair class to check for the presence of Granger causality.

4.4 Engle-Granger procedure

Let $\{y_t\}$ and $\{x_t\}$ be a pair of time series with $t \in H = \{1, ..., T\}$ the entire sample size of available history. Let $W = \{t_1, ..., t_R\} \subseteq T$ a smaller sample size of history.

Following Diamond [6], the Engle-Granger procedure can be described in three steps:

1. Obtain the fitted residual $\hat{e}_t$ and test for stationarity with the ADF test (reporting the significance), where:

$$y_t = b_0 + b_2 x_t + \epsilon_t, \tag{9}$$

$$\hat{e}_t = y_t - \hat{b}_2 x_t - \hat{b}_0. \tag{10}$$

   If the fitted residual is not stationary, then no long-run relationship exists and regression is spurious. Note here that we have a choice of the lookback period to use on which to test for stationarity. We could test on the entire horizon H or on a smaller sample size W. In a trading context we could use the smallest sample size that allows us to not reject the absence of stationarity and reserve subsequent data for out of sample backtesting, a concept we explore in further sections.

2. Plug the stationary fitted residual $\hat{e}_{t-1}$ from previous step, shifted, into error correction linear regression and confirm statistical significance of its coefficient $(1 - \alpha)$:

$$\Delta y_t = \phi \Delta x_t - (1 - \alpha)\hat{e}_{t-1}. \tag{11}$$

3. Fit to an Ornstein-Uhlenbeck process: the linear cointegrating combination produces a stationary and mean-reverting spread. Reversion speed θ and bounds calculated as $\frac{\sigma_{OU}}{\sqrt{2\theta}}$ where e_t follows the stochastic differential equation:

$$de_t = -\theta(e_t - \mu)dt + \sigma_{OU} dW_t, \tag{12}$$

   rewritten as:

$$e_{t+1} = C + Be_t + \epsilon_t. \tag{13}$$

   Once regression is estimated, we can solve for:

$$\theta = -\frac{\ln(B)}{\tau} = -252 \times \ln(B) \text{ (assuming daily bars)}, \tag{14}$$

$$\mu = \frac{C}{1 - B}, \tag{15}$$

$$\sigma_{OU} = \sqrt{\frac{SSE \times 2\theta\tau}{1 - \exp(-2\theta\tau)}}. \tag{16}$$

We implement the Engle-Granger procedure from first principles in our MyPair python class.

4.5 Johansen procedure for a VECM

With the Johansen procedure for a Vector Error Correction Model (VECM), the causal relationship between a pair of price time series can be examined as:

$$\Delta \boldsymbol{X}_t = \sum_{i=1}^{p-1} \boldsymbol{\Gamma}_i \Delta \boldsymbol{X}_{t-i} + \boldsymbol{\Pi} \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t \sim N(0, \Sigma), \tag{17}$$

where $\boldsymbol{X}_t$ is 2×1 vector each being $I(1)$ such that their differences are $I(0)$. $\boldsymbol{\Gamma}_i$ and $\boldsymbol{\Pi}$ are 2×2 coefficient matrices measuring the short- and long-run adjustments of the system to changes in $\boldsymbol{X}_t$. In the Johansen framework, the existence of cointegration is tested through the λ_{max} and λ_{trace} statistics which test the rank of $\boldsymbol{\Pi}$:

- If $\boldsymbol{\Pi}$ has full rank (i.e. 2), then all variables in $\boldsymbol{X_t}$ are $I(0)$ and we should develop a VAR model in price levels;
- If $rank(\boldsymbol{\Pi}) = 0$, the VECM is reduced to a VAR model in first differences;
- if $rank(\boldsymbol{\Pi}) = 1$, there exists one cointegrating vector and $\boldsymbol{\Pi} = \boldsymbol{\alpha}\boldsymbol{\beta}'$ where $\boldsymbol{\beta}'$ is the vector of cointegrating parameters and $\boldsymbol{\alpha}$ the vector of error correction coefficients measuring the speed of convergence to the long-run steady-state.

We use the Johansen procedure from off-the-shelf libraries (statsmodels vecm) in our MyPair python class. It allows us to consider additional lags in the error correction equation compared to E-G.

4.6 Kalman filtering

In some settings, it could make sense to reestimate regression parameters with the same frequency as price data, e.g. if price information arrives daily, we could reestimate the regression coefficients daily as well. The Kalman filter is a two-step prediction and correction estimator algorithm, described in Lubnau & Todorova [7] and Chan [3]. We are trying to predict the "state" of an object (in our case intercept b_0 and slope b_2) while knowing the previous state estimation, the actual state prediction (based on the model) and the actual measurement. The Kalman equations are:

$$\boldsymbol{x}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{B}\boldsymbol{u}_t + \boldsymbol{w}_{t-1}, \tag{18}$$

where:

$$\begin{aligned} \boldsymbol{x}_t &\text{ is the state we are trying to predict,} \\ \boldsymbol{x}_{t-1} &\text{ is the previous state,} \\ \boldsymbol{u}_t &\text{ is an optional command input,} \\ \boldsymbol{w}_{t-1} &\text{ is process noise.} \end{aligned}$$

The above relates the current state with the state at a previous time step plus an external optional control and process noise.

The second (measurement) equation relates the state with the measurement:

$$\boldsymbol{z}_t = \boldsymbol{H}\boldsymbol{x}_t + \boldsymbol{v}_t, \tag{19}$$

where:

$\boldsymbol{z}_t$ is the current measurement,
$\boldsymbol{H} = \boldsymbol{I}$ in the special case we are directly measuring the state,
$\boldsymbol{v}_t$ is the measurement noise.

The errors $\boldsymbol{v}$ and $\boldsymbol{w}$ are assumed Normal with mean 0 and covariances $\boldsymbol{Q}$ and $\boldsymbol{R}$ respectively. The iteration process includes the two following steps:

1. Time update:

$$\hat{\boldsymbol{x}}_t = \boldsymbol{A}\boldsymbol{x}_{t-1} + \boldsymbol{B}\boldsymbol{u}_t \text{ (state prediction)}, \tag{20}$$
$$\hat{\boldsymbol{P}}_t = \boldsymbol{A}\boldsymbol{P}_{t-1}\boldsymbol{A}^T + \boldsymbol{Q} \text{ (uncertainty prediction)}. \tag{21}$$

2. Measurement update:

$$\boldsymbol{K}_t = \boldsymbol{P}_t\boldsymbol{H}^T(\boldsymbol{H}\boldsymbol{P}_t\boldsymbol{H}^T + \boldsymbol{R})^{-1} \text{ (Kalman gain)}, \tag{22}$$
$$\boldsymbol{x}_t = \hat{\boldsymbol{x}}_t + \boldsymbol{K}_t(\boldsymbol{z}_t - \boldsymbol{H}\hat{\boldsymbol{x}}_t) \text{ (uncertainty prediction)}, \tag{23}$$
$$\boldsymbol{P}_t = (\boldsymbol{I} - \boldsymbol{K}_t\boldsymbol{H})\hat{\boldsymbol{P}}_t \text{ (uncertainty update)}. \tag{24}$$

We use a MyKalmanFilter class and encapsulate it in our MyPair class to compute a spread. The filter is recalculated every trading day when new futures prices become available. The current day's data is used to predict the hedge ratio used on the following day.

5 Python implementation

Our python code includes:

1. A LoadEnergyFutures file which loads data from text and csv files and makes various pairs (timespreads, cracks and locational pairs). Because our data goes so far back in time, we had to collate it from various files thus this function is somewhat involved. Note that each time a pair is instantiated, the various procedures below (Engle-Granger, Johansen, Kalman) are run for that pair;

2. A MyLR file which implements a linear regression function which captures the normal equations of ordinary least squares for cross-sectional data and associated estimation computations (AIC, BIC, t-stats, etc.). This function is encapsulated in a MyOLS class which attempts to mimic the statsmodel OLS class. Note our implementation is very basic and does not check for nonconstant variance, nonnormality or nonlinearity, which are important assumptions in OLS modeling;

3. A MyVAR class which implements a Vector Autoregression model leveraging our MyOLS class mimicking the statsmodel version. It is associated to a likelihood ratio test function which implements a test to detect the presence of Granger causality;

4. A MyKalmanFilter class which implements the Kalman filter algorithm;

5. A MyPair class which encapsulates all data processes for cleaning the price levels and running the Engle-Granger (E-G) procedure, Johansen procedure and Kalman filter algorithm, as well as getting processed outputs. Our MyPair class makes use of the classes above;

6. A Backtest file which implements entries and exits on the spread, as well as two backtest approaches:

   6.1. A first approach uses the entire horizon of available data to estimate the spread and run the backtest;

   6.2. A second approach slices the entire horizon in chunks and rolls a window to first estimate the spread then backtest on out-of-sample data to report trading results on out-of-sample data only. This approach is used for the spread obtained via E-G only.

7. A Utilities file implementing functions for plotting and running ADF tests.

8. A companion Jupyter notebook carrying the entire analysis workflow.

6 Estimation using full history

6.1 Cointegration results

6.1.1 Engle-Granger and Johansen procedures

Table 1 summarizes the cointegration estimates for both the Engle-Granger (E-G) and Johansen (Joh) procedures for timespreads under consideration using the full history of the sample. We can observe that:

1. Six timespreads have half-lives below 30 days (as determined by the E-G procedure) and are GO13, RB13, HO13, RB112, CL13 and NG13, i.e. typically the timespreads for the front months that see the most trading volumes. Those are good candidates on which to focus mean-reversion trading strategies. Other spreads are above 30 days and likely are not good candidates.

2. The Johansen procedure fails for RB13 (i.e. $rank(\mathbf{\Pi}) \neq 1$) but succeeds for all other pairs.

3. The t-statistics of both steps are deemed significant at 5% for all spreads for the E-G procedure. (Note we do not check the t-stat in the opposite direction as in practice for futures the spread works in one direction only for liquidity reasons. We would check it if we were working with another asset class e.g. equities.)

4. We find the presence of Granger causality (in both directions) in all pairs.

5. The difference between the spreads obtained via the E-G and the Johansen procedures quite small; as an example Figure 1 displays the spreads obtained via both procedures and their difference for LCO13. The difference is due to the inclusion of additional lags in the short-term correction equation via Johansen. We found this to be typical for all spreads considered. This is not the case with the spread obtained via Kalman filter as shown subsequently.

Results for cracks are shown in Table 2:

name	half_life	step1_ADF_tstat	step2_ec_tstat	b_2 (E-G)	b_0 (E-G)	θ	μ	σ_{OU}	GC 1-2	GC 2-1	b_2 (Joh)	b_0 (Joh)	lag (Joh)
GO13	12.32	-9.51	-12.20	-1.03	-0.01	14.17	0.000	0.042	1	1	-1.02	0.01	4
RB13	21.05	-7.26	-8.67	-1.05	-0.05	8.30	0.000	0.057	1	1	0.00	0.00	1
HO13	23.86	-11.19	-11.42	-0.97	0.00	7.32	0.000	0.078	1	1	-0.97	0.00	2
RB112	25.45	-7.00	-7.70	-1.13	-0.16	6.86	0.000	0.075	1	1	-1.14	0.16	1
CL13	27.49	-10.67	-10.85	-0.99	0.01	6.35	0.000	0.083	1	1	-0.98	-0.01	4
LCO13	29.43	-9.12	-10.56	-0.99	0.01	5.93	0.000	0.065	1	1	-0.99	-0.01	3
NG13	29.65	-9.46	-9.59	-1.60	-0.03	5.89	0.005	1.419	1	1	-1.59	0.00	0
LCO36	31.55	-10.14	-10.11	-0.99	0.01	5.54	0.000	0.070	1	1	-0.99	-0.01	6
GO16	41.88	-5.84	-6.81	-1.03	-0.01	4.17	0.001	0.086	1	1	-1.02	0.00	2
LCO16	43.28	-8.48	-8.80	-0.98	0.02	4.04	0.000	0.129	1	1	-0.98	-0.02	6
RB16	45.21	-5.26	-5.76	-1.08	-0.08	3.86	0.000	0.100	1	1	-1.09	0.08	1
CL16	49.11	-8.65	-8.19	-0.97	0.02	3.56	0.000	0.147	1	1	-0.97	-0.03	4
HO16	49.42	-8.27	-7.98	-0.93	0.01	3.53	0.000	0.141	1	1	-0.93	-0.01	0
NG16	52.48	-7.31	-7.37	-1.80	0.04	3.33	0.009	2.434	1	1	-1.77	-0.10	0
NG36	55.53	-7.23	-7.44	-1.13	0.03	3.15	0.000	1.150	1	1	-1.13	-0.03	0
NG112	58.91	-6.88	-7.79	-1.32	0.02	2.96	-0.002	2.850	1	1	-1.31	-0.03	0
RB36	61.28	-5.01	-4.93	-1.04	-0.03	2.85	0.000	0.071	1	1	-1.06	0.05	0
HO112	65.50	-7.13	-7.05	-0.94	0.01	2.67	0.000	0.225	1	1	-0.94	-0.01	1
CL112	74.46	-7.07	-6.82	-0.96	0.04	2.35	0.000	0.224	1	1	-0.95	-0.05	4
CL36	75.55	-6.91	-6.45	-0.99	0.01	2.31	0.000	0.071	1	1	-0.98	-0.02	3
GO112	84.78	-4.44	-5.12	-1.00	0.02	2.06	0.002	0.142	1	1	-1.01	0.00	1
HO36	87.41	-6.20	-5.63	-0.96	0.01	2.00	0.000	0.088	1	1	-0.96	-0.01	1
GO36	112.38	-4.32	-4.70	-1.01	0.00	1.55	0.001	0.048	1	1	-1.01	0.00	1

Table 1: Cointegration estimates for timespreads using full history.

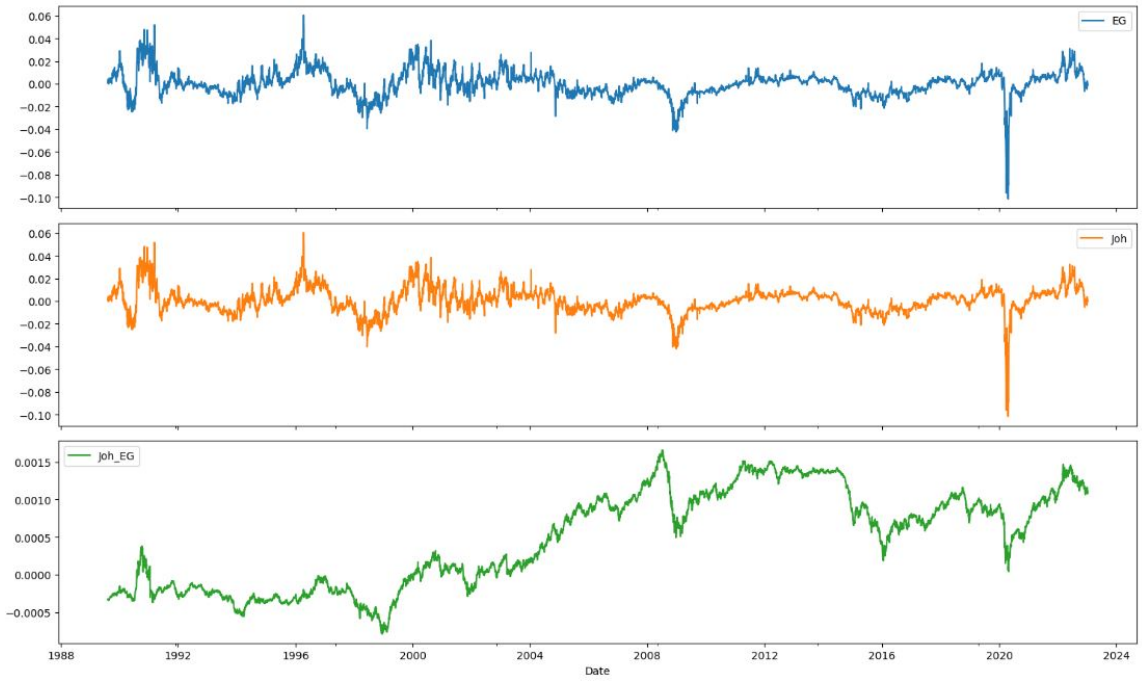

Figure 1: Spreads obtained via E-G and Johansen procedures and their difference for LCO13.

1. LCOGO6, LCOGO1, LCOGO3 and CLRB1 all have half-lives below 30 and are good candidates for mean-reverting strategies. This is similar to Alizadeh & Nomikos [1] and Cummins & Bucca [4] who find that spreads which include GO in one leg typically perform better (shortest half-lives, higher Sharpe ratios).
2. The Johansen procedure fails for these pairs as well as CLRBC3.
3. The t-statistics of both steps are deemed significant at 5% for all pairs.
4. We find the presence of Granger causality (in both directions) in all pairs.

Results for locational spreads are shown in Table 3:

1. HO2GO1, HO4GO3, HO7GO6 and CL1LCO2 all have half-lives below 30 and are good candidates for mean-reverting strategies.
2. The Johansen procedure succeeds for all these pairs.
3. The t-statistics of both steps are deemed significant at 5% for all pairs.
4. We find the presence of Granger causality (in both directions) in all pairs.

6.1.2 Kalman filter procedure

The spreads obtained via Kalman filtering appear more volatile than those obtained via E-G or Johansen approaches, as exhibited in Figure 2, likely due to the more frequent reestimation. As an illustration, Figure 3 displays the time-varying b_2 obtained via Kalman filtering for GO13, which seems to converge towards the E-G b_2 of Table 1, and b_0 which converges towards 0 as expected.

6.2 Trading results

Following Diamond [5], for the E-G spread the trading rules are:

1. When the spread crosses the upper level Z from below we enter a short position, expecting it to return to its mean. We exit the short position when it crosses back to its mean from above, realizing a profit of $\mid Z - \mu \mid$;
2. When the spread crosses the lower level $-Z$ from above we enter a long position, expecting it to return to its mean. We exit the long position when it crosses back to its mean from below, realizing a profit of $\mid Z - \mu \mid$.

Following Lubnau & Todorova [7], for the Kalman spread, the trading rules are modified as follows:

1. Define the z-score indicator $z_t = (s_t - MA_t)\sigma_t^{-1}$, where MA_t and σ_t denote simple moving average and rolling standard deviation defined over same length (typically 20 days);
2. When z_t crosses the upper level 2 from below we enter a short position, expecting it to return to its mean. We exit the short position when it crosses back to 0 from above;
3. When the spread crosses the lower level -2 from above we enter a long position, expecting it to return to its mean. We exit the long position when it crosses back to 0 from below.

name	half_life	step1_ADF_tstat	step2_ec_tstat	b_2 (E-G)	b_0 (E-G)	θ	μ	σ_{OU}	GC 1-2	GC 2-1	b_2 (Joh)	b_0 (Joh)	lag (Joh)
LCOGO6	15.64	-6.82	-8.53	-1.03	-0.04	11.17	0.00	0.07	1	1	0.00	0.00	5
LCOGO1	17.09	-7.83	-6.79	-1.00	-0.03	10.22	0.00	0.10	1	1	0.00	0.00	4
LCOGO3	18.03	-6.60	-7.65	-1.02	-0.03	9.69	0.00	0.08	1	1	0.00	0.00	3
CLRB1	29.46	-7.21	-7.16	-1.05	-0.05	5.93	0.00	0.11	1	1	0.00	0.00	5
CLHO1	40.19	-8.80	-7.87	-0.96	0.03	4.35	0.00	0.19	1	1	-0.96	-0.04	7
CLRB3	60.67	-4.98	-5.19	-1.05	-0.05	2.88	0.00	0.10	1	1	0.00	0.00	5
CLHO3	76.49	-5.71	-6.28	-0.99	0.02	2.28	0.00	0.16	1	1	-0.98	-0.03	1
CLRB6	80.35	-4.13	-4.08	-1.07	-0.07	2.17	0.00	0.09	1	1	-1.16	0.15	1
CLHO6	100.41	-4.75	-5.50	-1.00	0.02	1.74	0.00	0.14	1	1	-0.99	-0.02	1

Table 2: Cointegration estimates for cracks using full history.

name	half_life	step1_ADF_tstat	step2_ec_tstat	b_2 (E-G)	b_0 (E-G)	θ	μ	σ_{OU}	GC 1-2	GC 2-1	b_2 (Joh)	b_0 (Joh)	lag (Joh)
HO2GO1	5.21	-13.21	-14.36	-0.94	0.06	33.50	0.00	0.05	1	1	-0.95	-0.05	3
HO4GO3	5.71	-11.15	-14.70	-0.96	0.05	30.57	0.00	0.04	1	1	-0.96	-0.04	3
HO7GO6	7.06	-9.76	-12.69	-0.96	0.04	24.73	0.00	0.04	1	1	-0.97	-0.03	4
CL1LCO2	29.60	-9.57	-10.41	-0.88	0.12	5.90	0.00	0.13	1	1	-0.87	-0.12	4
CL3LCO4	34.24	-7.63	-9.68	-0.88	0.12	5.10	0.00	0.10	1	1	-0.88	-0.12	4

Table 3: Cointegration estimates for locational spreads using full history.

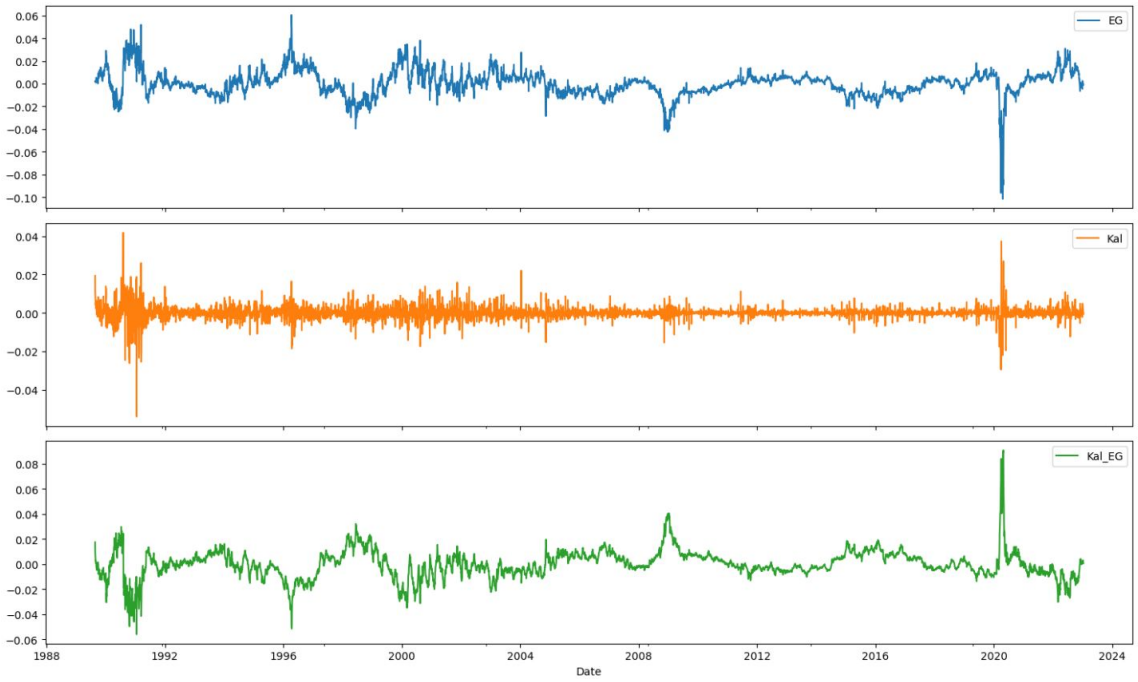

Figure 2: Spreads obtained via E-G and Kalman procedures and their difference for LCO13 (first 10d cropped).

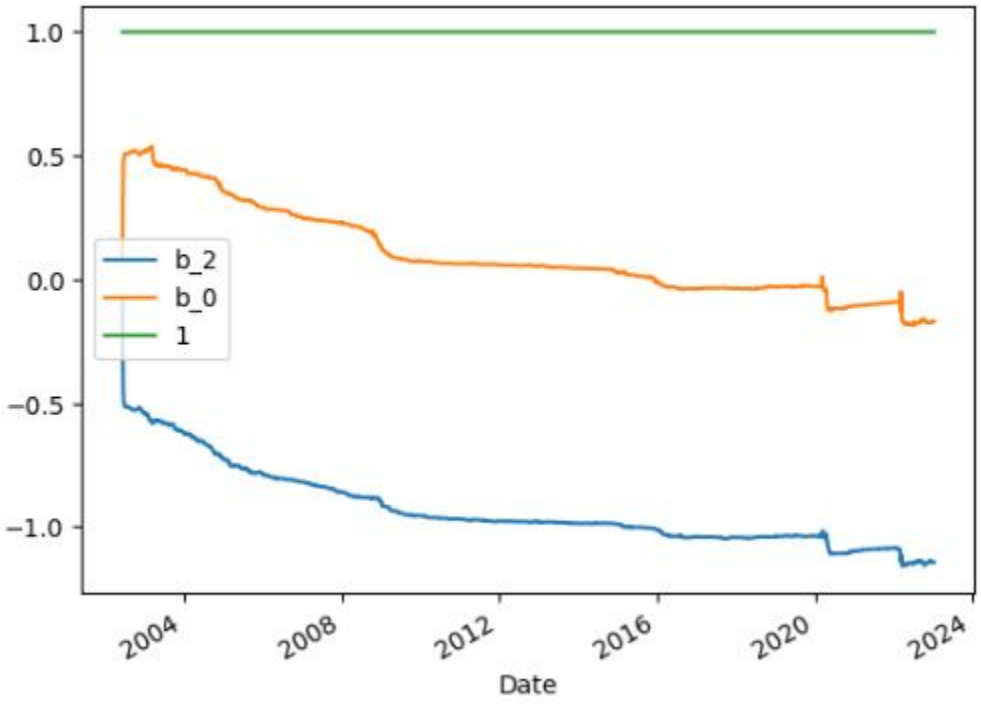

Figure 3: Coefficients in the cointegration vector obtained via Kalman filtering for GO13.

name	Avg annual return	Sharpe	max drawdown
HO4GO3	7.1%	1.12	-4.5%
LCOGO1	6.7%	0.8	-12.4%
LCOGO3	5.6%	0.82	-8.9%
HO2GO1	5.5%	0.95	-4.0%
RB112	5.3%	0.96	-10.0%
LCOGO6	5.3%	0.86	-7.4%
RB13	4.9%	1.16	-3.9%
CL1LCO2	4.0%	0.6	-26.5%
GO13	2.3%	0.71	-7.0%
HO13	2.0%	0.57	-8.2%
CL13	1.6%	0.41	-18.0%

Table 4: Trading results for selected E-G spreads using full history.

We refer the reader to the companion Jupyter notebook, where a search procedure for Z is conducted for the five timespreads with shortest half-lives. Of particular interest is HO13, for which we find $Z = 0.2$ to maximize Sharpe ratio, giving a Sharpe ratio of 0.57 and maximum drawdown of -8.2%. We plot entries and exits and display backtest tear sheets in Figures 4 and 5.

Trading results for 11 selected E-G spreads are displayed in Table 4, noting that spreads with GO in one leg provide the highest average annual return, a result already known in the literature and also experienced in our practice.

Trading results for 11 selected Kalman spreads are displayed in Table 5, we note that results are typically inferior to those obtained via E-G (Table 4). This could be due to the higher number of trades or the slower convergence of the cointegration vector obtained via Kalman. (This is not always the case though, e.g. CL13 in the Kalman setting yields better results than in the E-G setting.) This could perhaps be remediated by the removal of an initial "warm-up" period in the Kalman filter backtest (we leave this to future research). Interestingly, our results are in the same range as those obtained by Lubnau & Todorova [7] (keep in mind they report average 5-yr results over 1989-2014 including synthetic data while we report the 1-yr avg over 1989-present):

- For CL13 they report an avg annual return of 5.3% and Sharpe of 2.7; we have 2.5% and 0.83.
- For HO13 they report an avg annual return of 0.8% and Sharpe of 0.72; we have 3.1% and 0.97.
- For RB13 they report an avg annual return of 3.5% and Sharpe of 0.97; we have 3.4% and 0.92.

7 Rolling estimation

As mentioned in Diamond [6] (Appendix B, p. 25), practitioners advise estimating the cointegration using one year of historical data then trading the estimate for a six-month period. For our futures contract, we found that using a 3-year estimation window was necessary to accept the cointegration relationship at 5% significance in the E-G procedure. We conjecture (without proof) that the three year period is linked to the seasonality of oil inventories.

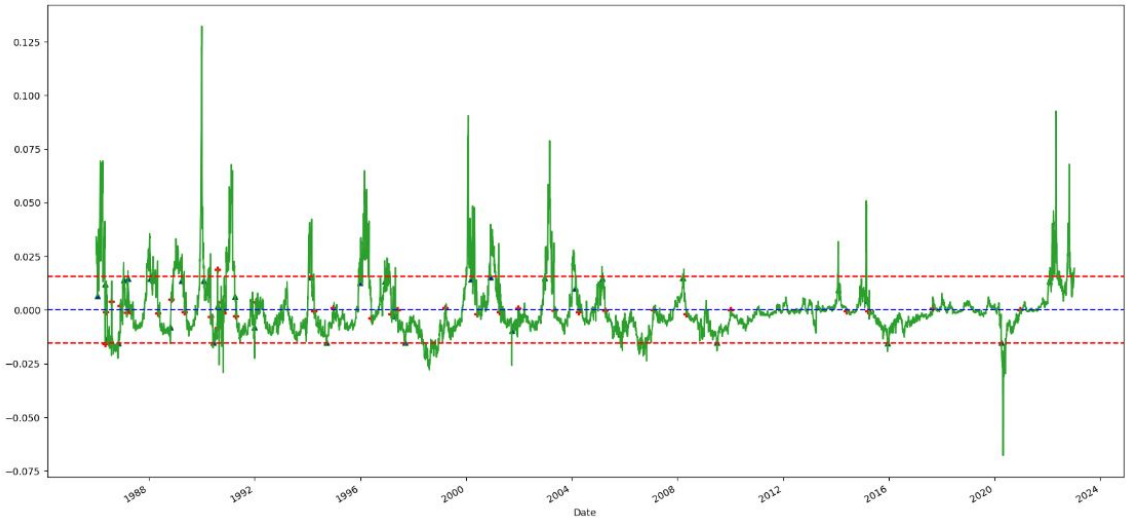

Figure 4: HO13 spread entries (blue triangles) and exits (red crosses) with Z = 0.2.

name	Avg annual return	Sharpe	max drawdown
HO4GO3	0.1%	0.05	-13.2%
LCOGO1	2.4%	0.46	-9.9%
LCOGO3	0.6%	0.18	-6.6%
HO2GO1	-0.4%	-0.14	-19.3%
RB112	5.4%	1.22	-7.9%
LCOGO6	4.2%	0.08	-8.4%
RB13	3.4%	0.92	-9.9%
CL1LCO2	3.7%	0.84	-15.3%
GO13	-0.6%	-0.13	-31.0%
HO13	3.1%	0.97	-5.2%
CL13	2.5%	0.83	-7.5%

Table 5: Trading results for selected Kalman spreads.

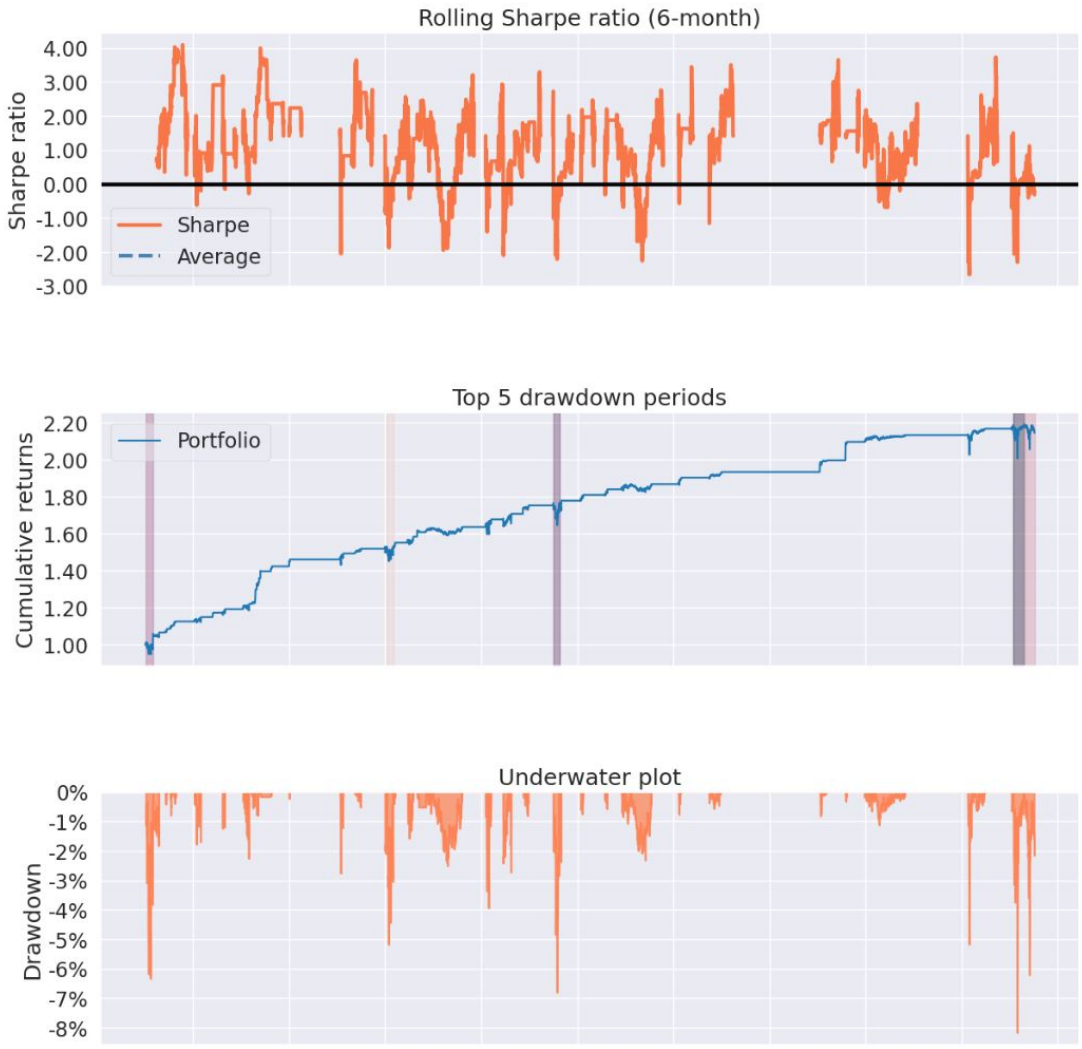

Figure 5: HO13 spread rolling Sharpe ratio and drawdown.

Additionally there is the problem of data snooping: when we use the entire history to estimate the spread then backtest on the entire history again, we are using data that we would not have had in live trading conditions.

That is why we devise a backtesting approach whereby we slice the entire horizon in chunks and roll a 3-year window to first estimate the spread then subsequently backtest on 4 months of out-of-sample data; we then stitch together out-of-sample-only trading results over the entire horizon (less the initial 3 years). Note that 3-years and 4-months are parameters of the rolling estimation/out-of-sample backtesting procedure and can be modified as needed.

We report below results for one timespread, one crack and one locational spread of particular interest:

1. Timespread GO13: Figure 6 shows the normalized price levels, t-stats of each steps of the E-G procedure, time-varying b_2 ratio and resulting out-of-sample spread from the rolling estimation. Observe the extreme volatility in the Spring of 2022 during the Russia/Ukraine conflict which resulted in "out-of-bounds" spread. Note that the whole-horizon estimation results in constant $b_2 = -1.02$ (Table 1) whereas in the rolling estimation b_2 oscillates above and below -1. Figure 7 displays the resulting time-varying Sharpe, drawdowns and underwater plots. Overall Sharpe is 0.61 and max drawdown -49.4% which is quite large and would need to be reduced via risk management (stop losses).

2. Crack LCOGO3: Figure 8 shows the normalized price levels, t-stats of each steps of the E-G procedure, time-varying b_2 ratio and resulting out-of-sample spread from the rolling estimation. The volatility in the Spring of 2022 during the Russia/Ukraine conflict is a bit less than in the timespread above. Note that the whole-horizon estimation results in constant $b_2 = -1.02$ (Table 2) whereas in the rolling estimation b_2 oscillates above and below -1. Figure 9 displays the resulting time-varying Sharpe, drawdowns and underwater plots. Overall Sharpe is 0.68 and max drawdown -24.1%. Note that there is an extended period of time during which the spread is within bounds and no trade is effectuated. This is likely associated to a period of weak refinery margins and refinery overcapacity.

3. Locational spread HO4GO3: Figure 10 shows the normalized price levels, t-stats of each steps of the E-G procedure, time-varying b_2 ratio and resulting out-of-sample spread from the rolling estimation. Note that the whole-horizon estimation results in constant $b_2 = -0.96$ (Table 3) whereas in the rolling estimation b_2 oscillates above and below -1. Figure 11 displays the resulting time-varying Sharpe, drawdowns and underwater plots. Overall Sharpe is 0.41 and max drawdown -58.5%. Note that there is an extended period of time during which the spread is within bounds and no trade is effectuated. Also note that there seems to be a trend in the spread which the E-G procedure (as coded) fails to capture; alternatively we would need to build a spread using the Johansen procedure with a trend in the cointegration relationship. The trend could be the effect of the Renewable Volume Obligation (RVO) in the United States which has increased in importance as a share of the NYMEX Heating Oil contract in recent memory. It is interesting for this environmental regulation to manifest itself here and we leave this to future research.

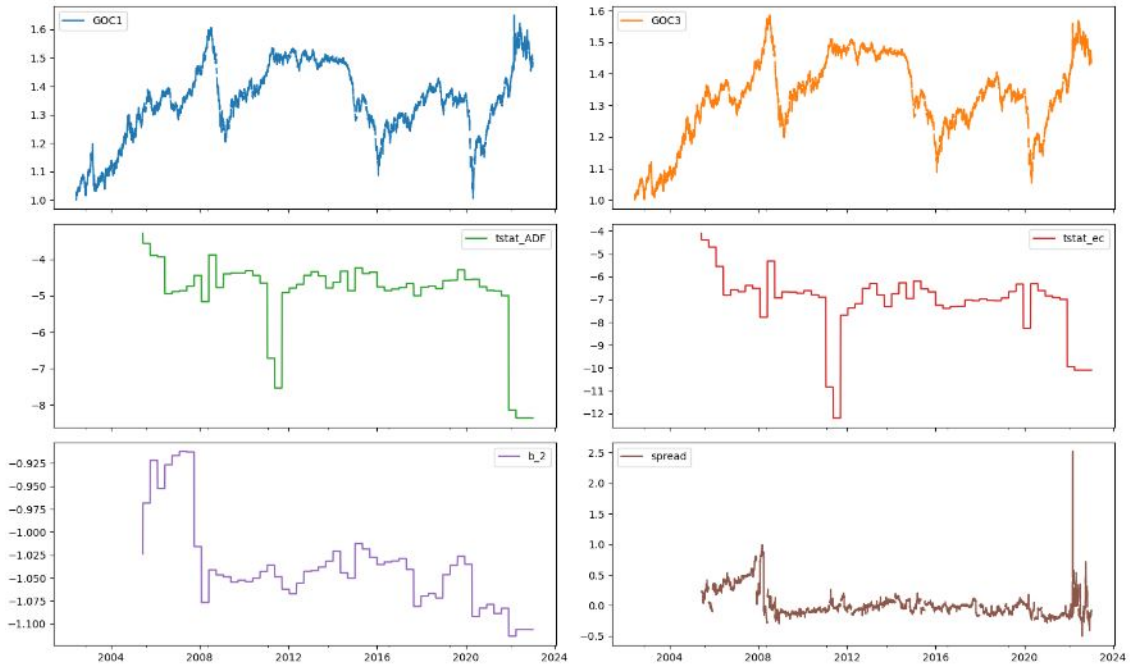

Figure 6: GO13 levels, t-stats, hedge ratio and out-of-sample spread from rolling estimation.

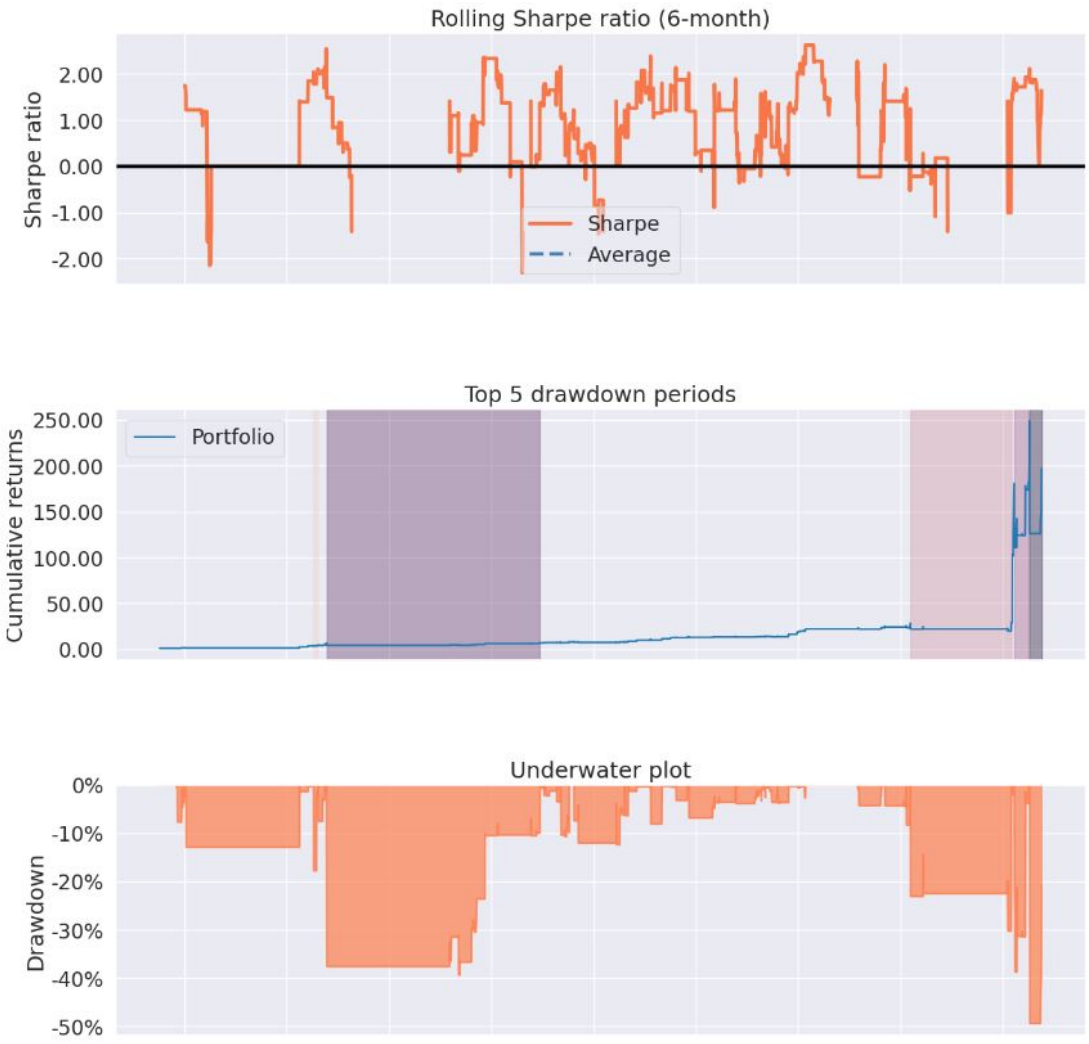

Figure 7: GO13 time-varying Sharpe, drawdowns from rolling estimation.

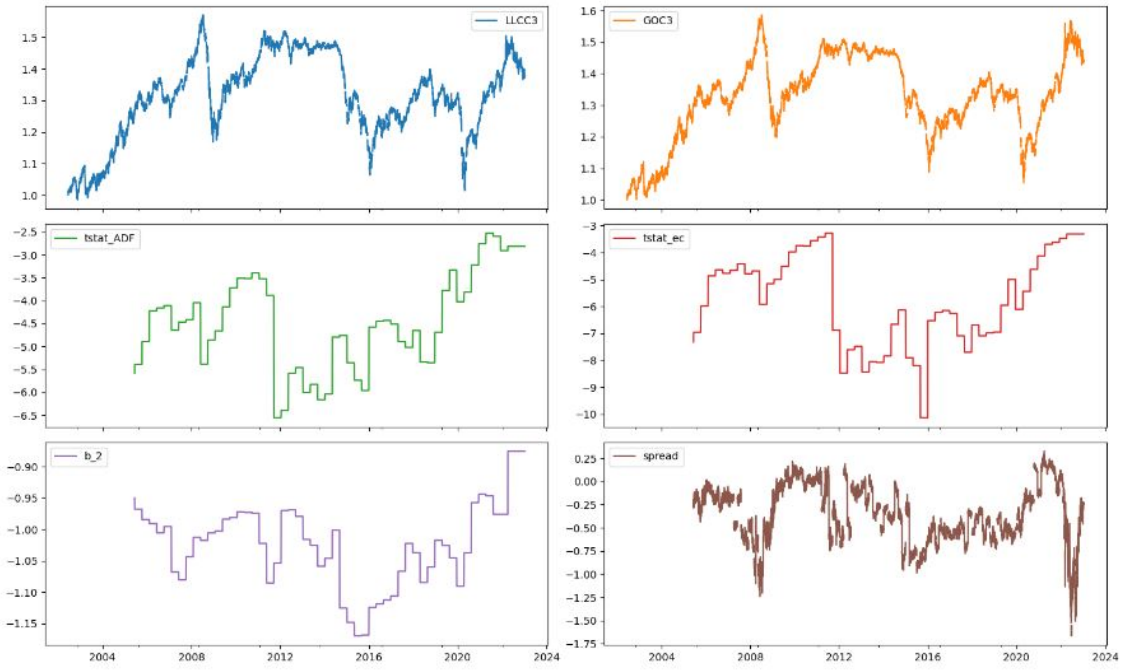

Figure 8: LCOGO3 levels, t-stats, hedge ratio and out-of-sample spread from rolling estimation.

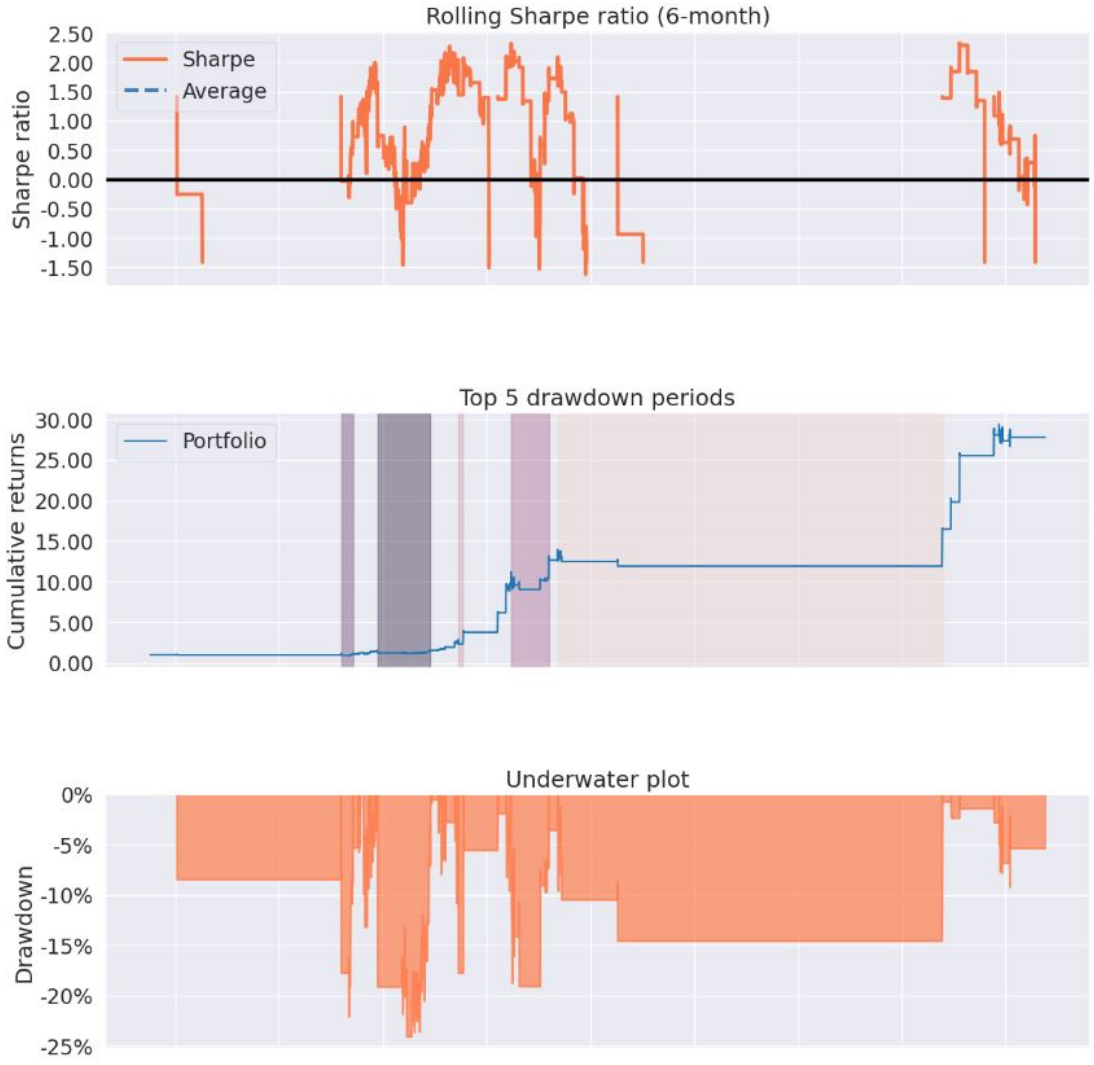

Figure 9: LCOGO3 time-varying Sharpe, drawdowns from rolling estimation.

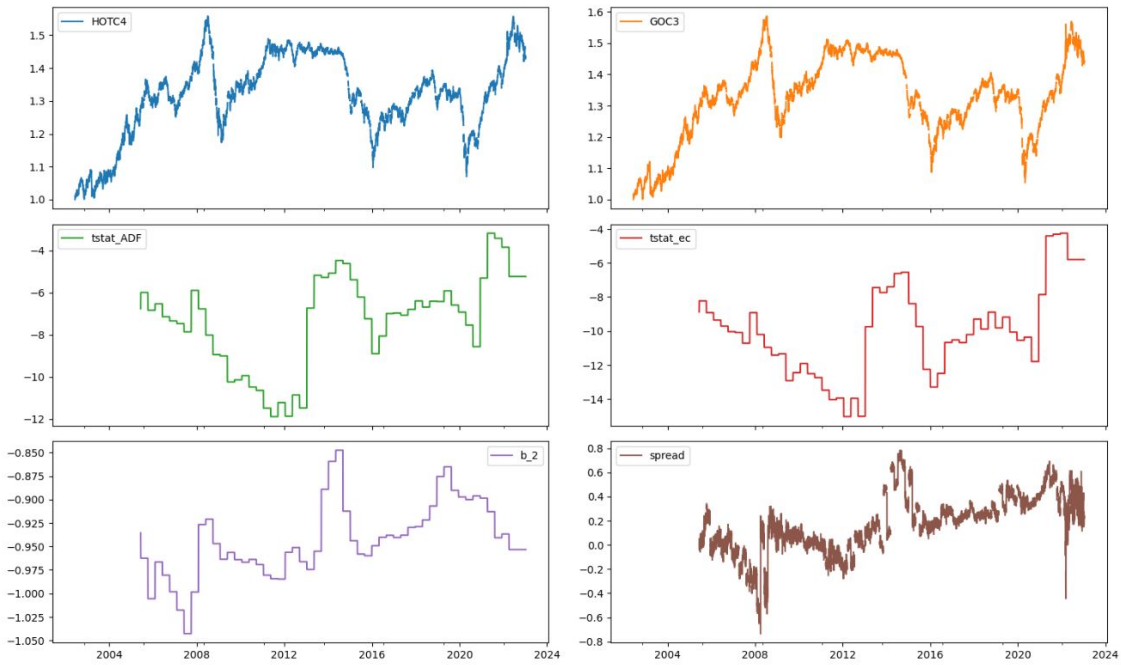

Figure 10: HO4GO3 levels, t-stats, hedge ratio and out-of-sample spread from rolling estimation.

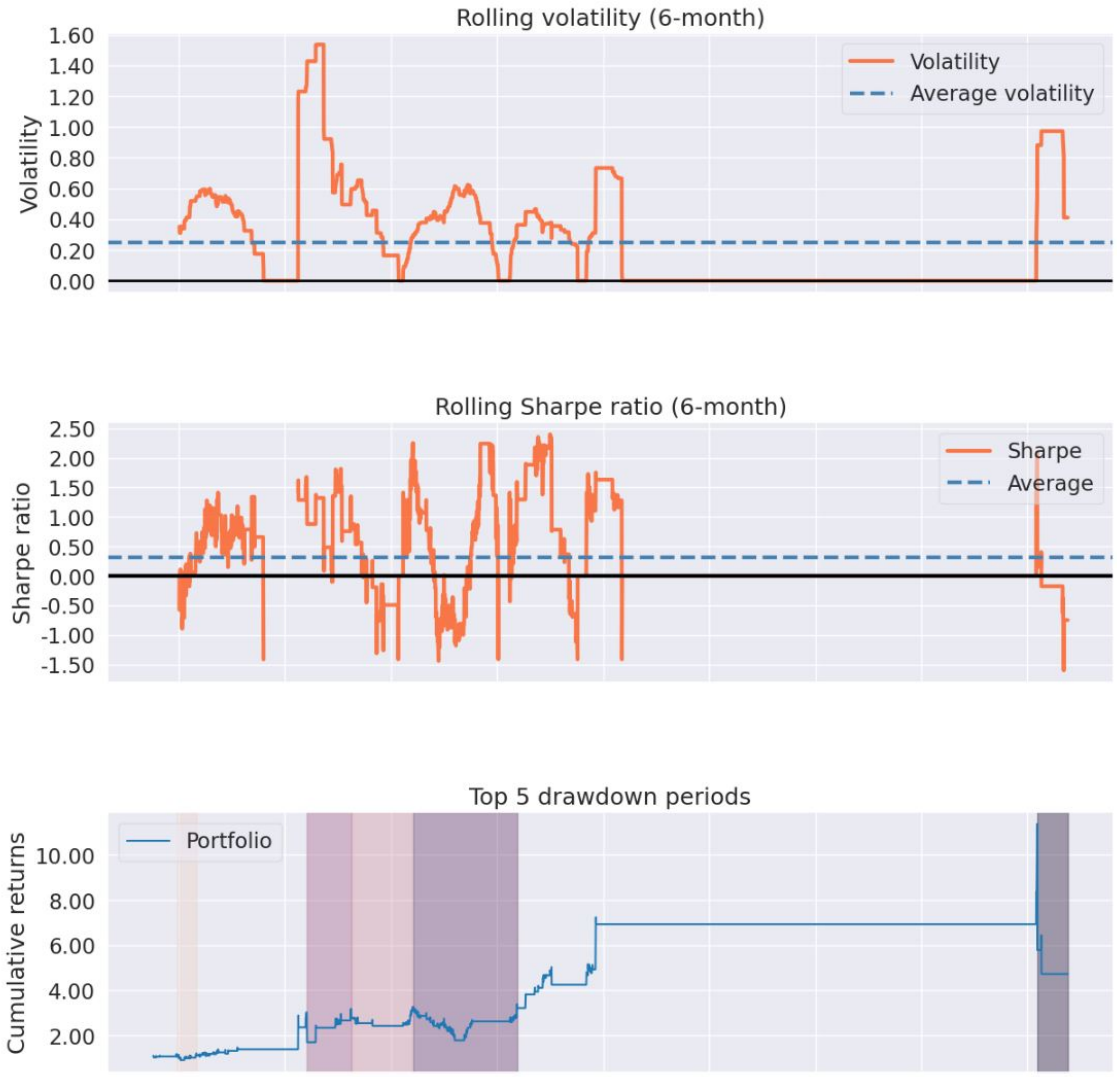

Figure 11: HO4GO3 time-varying Sharpe, drawdowns from rolling estimation.

8 Conclusion

Financial econometrics is a deep and fascinating subject and fortunes have been made with statistical arbitrage trading strategies. In this report, we design a pairs trading strategy for oil futures. We implement regression estimation in matrix form, a VAR function, a Granger causality test, the Engle-Granger procedure from first principles, the Johansen procedure from off-the-shelf libraries, the Kalman filter from first principles, and a backtest engine (in two flavors: whole horizon and rolling estimation-window/out-of-sample backtest) in the python programming language. This allows us to study calendar spreads, intercommodity spreads and locational spreads among a universe of 37 pairs indexed over 1989-present.

We find that the timespreads with half-lives below 30 days are: GO13, RB13, HO13, RB112, CL13, LCO13 and NG13. The cracks with half-lives below 30 days are: LCOGO6, LCOGO1, LCOGO3 and CLRB1. The locational with half-lives below 30 days are: HO2GO1, HO4GO3, HO7GO6 and CL1LCO2.

Using the Engle-Granger spread, we find that spreads yielding average annual returns above 5% include HO4GO3 (7.1%), LCOGO1 (6.7%), LCOGO3 (5.6%), HO2GO1 (5.5%), RB112 (5. %) and LCOGO6 (5.3%). Note the important of Gasoil (GO) in one of the legs. We would recommend focusing on these for a practical trading implementation.

Directions of future research include:

1. The consideration of transaction costs, slippage, volume and liquidity for a more realistic trading performance evaluation;

2. The Kalman filter algorithm has parameters for R and Q, it would be interesting to vary these and test the sensitivity of the algorithm to these parameters;

3. Bertram [2] provides anlytical formulae for optimal trading under an Ornstein-Uhlenbeck process. It would be interesting to implement these within our strategy;

4. Use of synthetic data (via a stationary bootstrap algorithm on returns) to detect the presence of potentially false trading discoveries.

5. The construction of an optimal portfolio of energy futures spreads.

References

[1] A. Alizadeh and N. Nomikos, *Performance of statistical arbitrage in petroleum futures markets*, Journal of Energy Markets, (2008).

[2] W. bertram, *Analytic solutions for optimal statistical arbitrage trading*, Physica A, (2010).

[3] E. P. Chan, *Algorithmic Trading*, Wiley, 2013.

[4] M. Cummins and A. Bucca, *Quantitative spread trading on crude oil and refined products markets*, Quantitative Finance, (2012).

[5] R. Diamond, *Modeling long-term relationships in time series (cqf lecture)*.

[6] ——, *Learning and trusting cointegration in statistical arbitrage*, Wilmott, (2014).

[7] T. Lubnau and N. Todorova, *Trading on mean-reversion in energy futures markets*, Energy Economics, (2015).

[8] T. Nakajima, *Expectations for statistical arbitrage in energy futures markets*, Journal of Risk and Financial Management, (2019).

[9] D. Ruppert and D. Matteson, *Statistics and Data Analysis for Financial Engineering*, Springer, 2015.