Harry Harutyun Grigoryan (hg2592)

Problem: Non-Negative Matrix Factorization

Futures contracts are one of the liquid assets in the financial markets. Those contracts may be set on various asset classes to be transacted, such as commodities or financial instruments. We will focus on commodity futures such as gold for the context. As known, a commodity may have two prices, such as a spot price and a futures price. The spot price is the current price of a commodity, and the futures price is the price set in the futures contract. Those prices are associated with each other throughout the lifetime of a futures contract. On the other hand, the intraday trading volume of a security is the total amount of traded contracts distributed over the day. In this report, we will examine futures price and intraday traded volume with the Non-negative Matrix Factorization (NMF) technique and compare the results to that of PCA.

Consider the gold spot price of the futures contract (ticker: GC=F) and several futures contracts with different maturity dates on gold. The contracts are set for various months in the years 2023, 2024, 2025, and 2026:

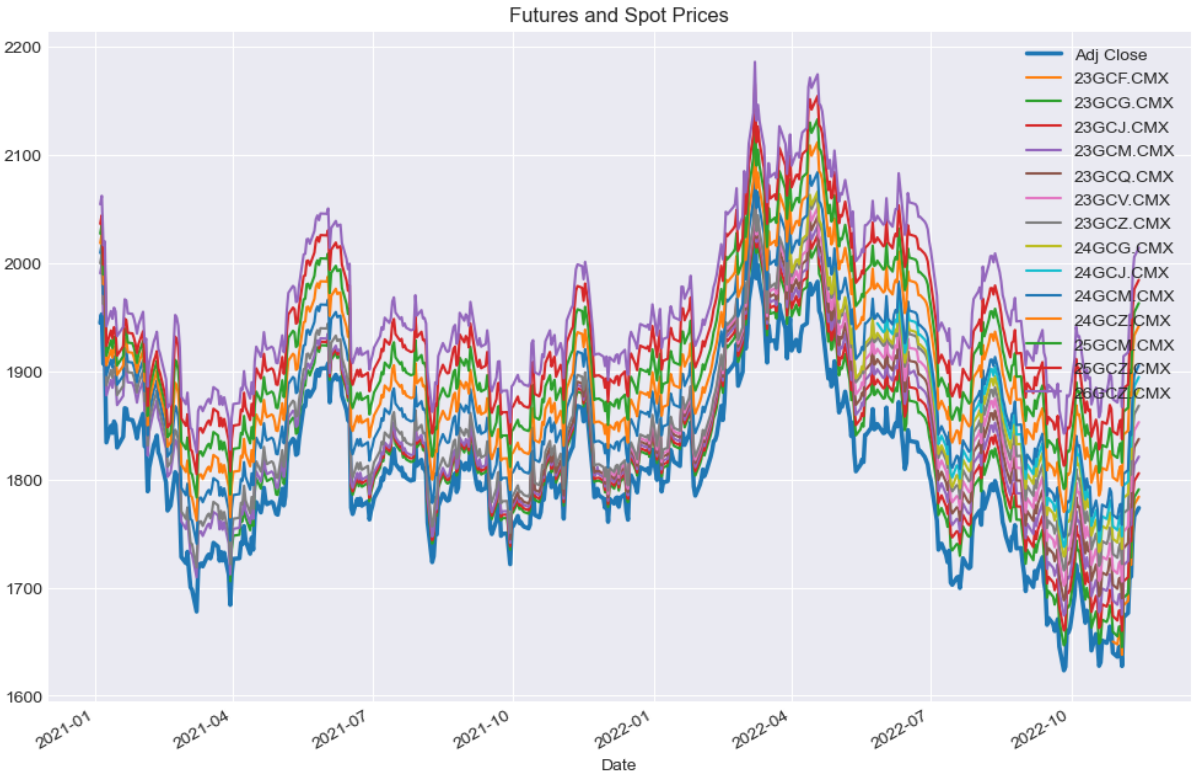
'GCF23.CMX', 'GCG23.CMX', 'GCJ23.CMX', 'GCM23.CMX', 'GCQ23.CMX','GCV23.CMX', 'GCZ23.CMX', 'GCG24.CMX', 'GCJ24.CMX', 'GCM24.CMX', 'GCZ24.CMX', 'GCM25.CMX', 'GCZ25.CMX', 'GCZ26.CMX'

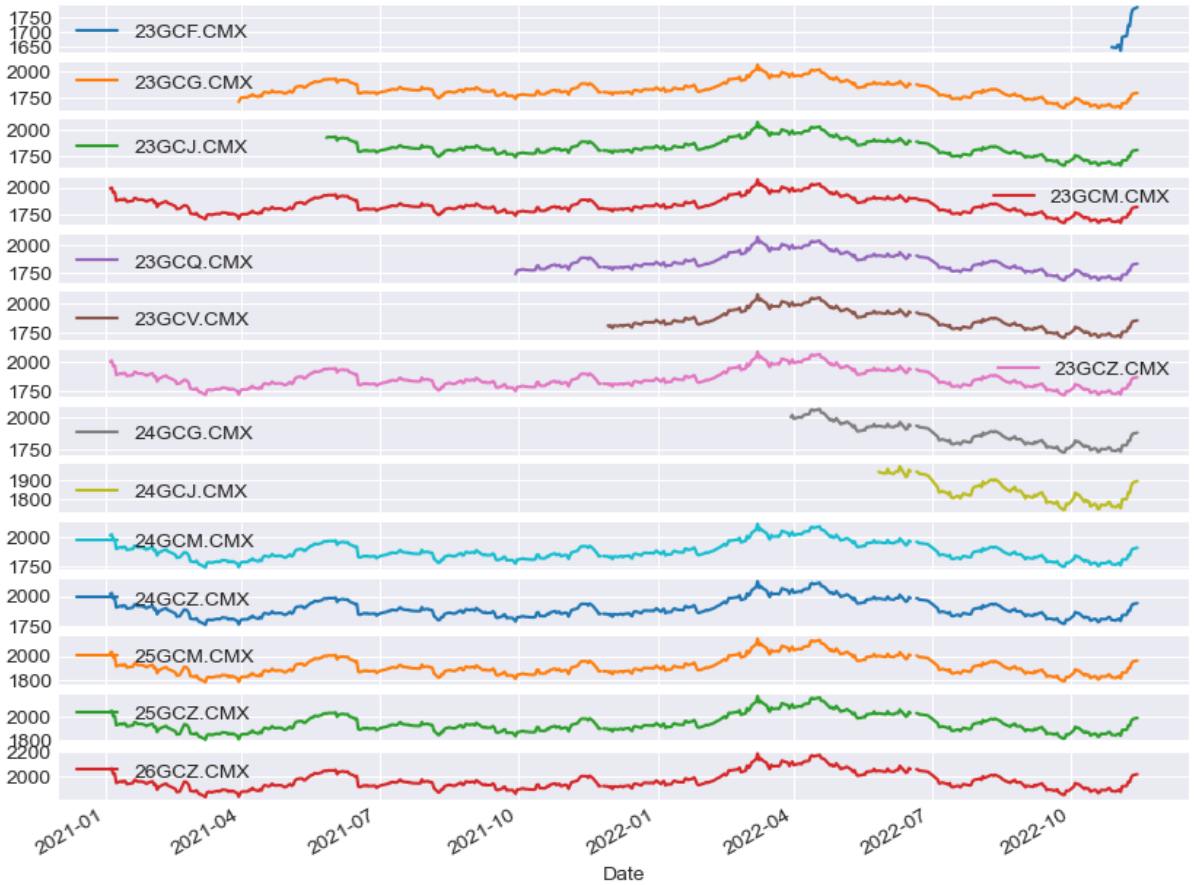
Example: **GS**: gold ticker, **F**: maturity month, **23**: year, **.CMX**: exchange

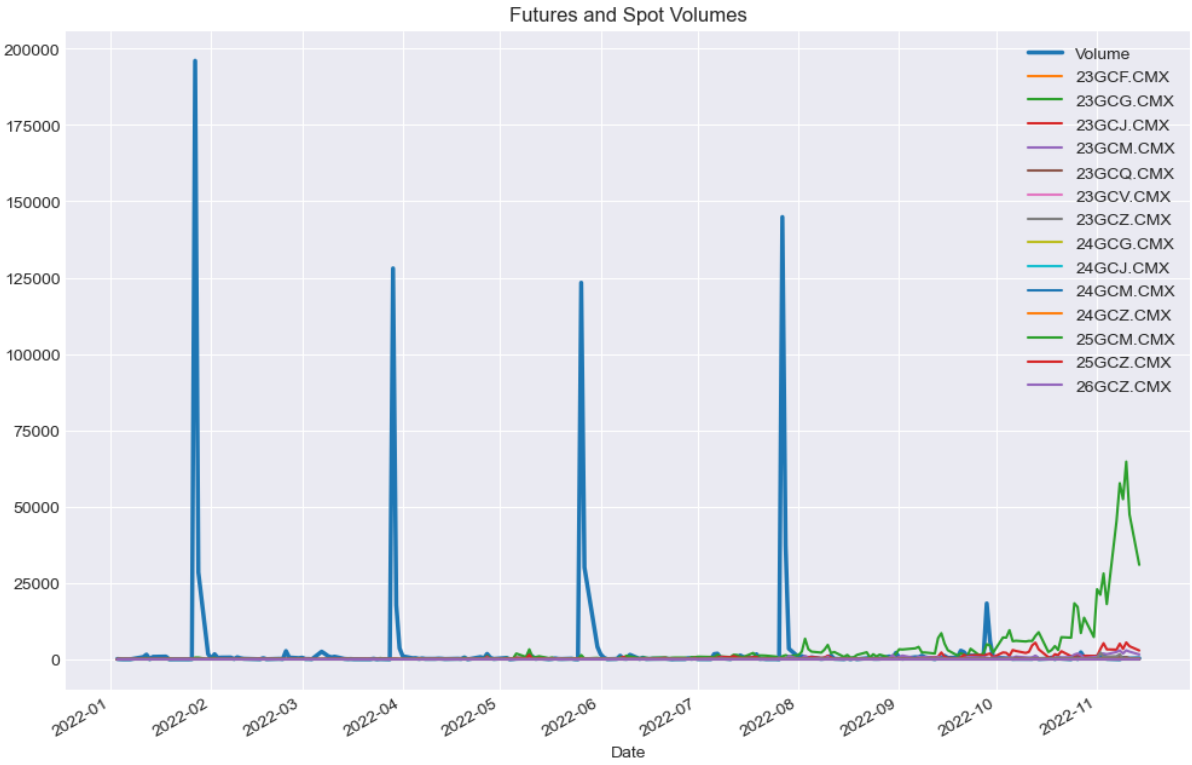
Now let us examine the price and traded volume daily. In the graphs below, the first one is the plot for spot and future contract prices. The bold blue line at the bottom is the spot price found in the legend. Usually, the spot price is less than the futures price but otherwise is also true. This report will not go deep into future price dynamics for now.

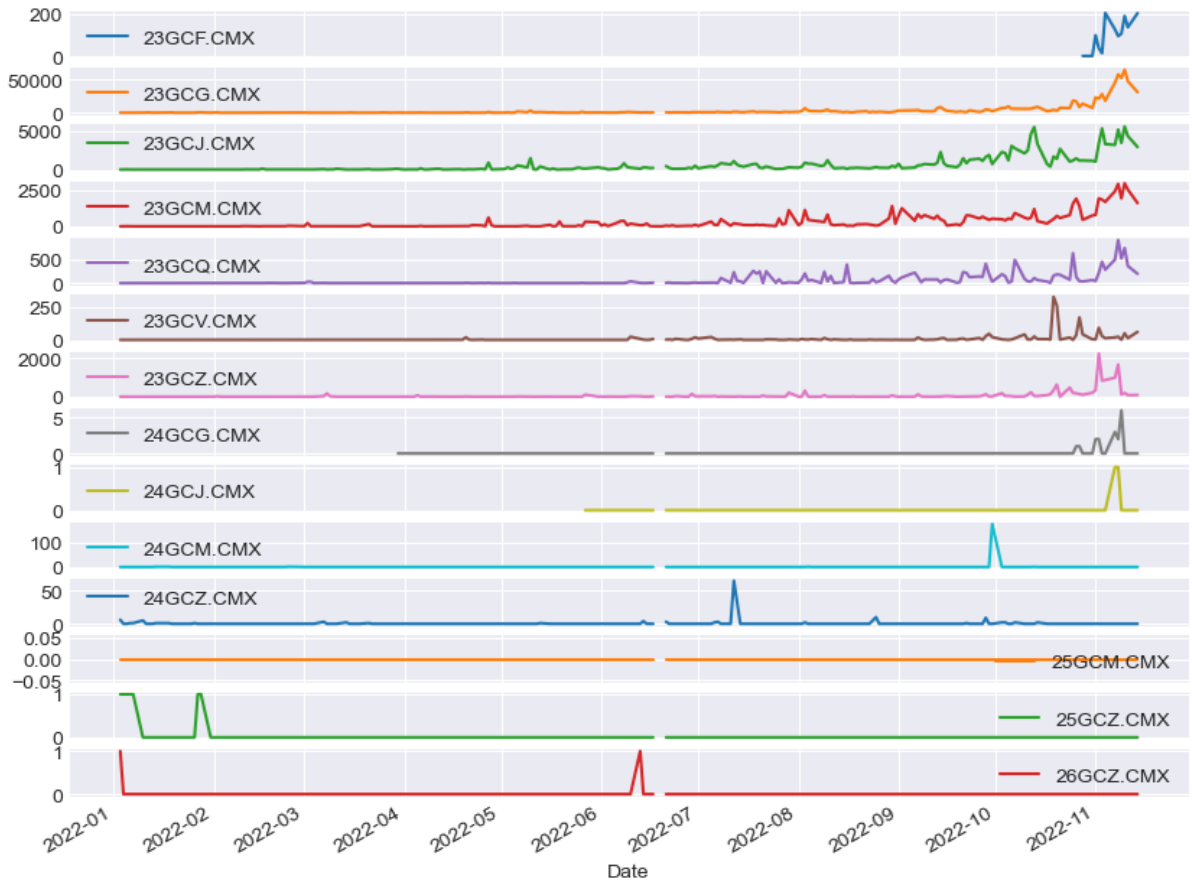
As we have the underlying component the same for all (i.e., gold), the pattern of price movements is reminiscent but on different scales because of each instrument's maturity factor. However, it is different in the case of traded volume. On the next page is the graphs for traded volumes. The spikes recorded for the spot (i.e., futures combined shown in blue) note the maturity dates as traders rush to close the positions before maturity. Examining the volume plots, we see how the maturity affects the traded volumes, and the concept of days-to-maturity arises. The traded volume increases towards the instrument's maturity as it should for the above-stated reason mostly.

The following sections will examine the price and volume dynamics with NMF and PCA matrix factorization techniques utilizing sklearn and NIMFA packages.


Futures and Spot Prices
Adj Close
23GCF.CMX
23GCG.CMX
23GCJ.CMX
23GCM.CMX
23GCQ.CMX
23GCV.CMX
23GCZ.CMX
24GCG.CMX
24GCJ.CMX
24GCM.CMX
24GCZ.CMX
25GCM.CMX
25GCZ.CMX
26GCZ.CMX
2200
2100
2000
1900
1800
1700
1600
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
2022-07
2022-10
Date


23GCF.CMX
23GCG.CMX
23GCJ.CMX
23GCM.CMX
23GCQ.CMX
23GCV.CMX
23GCZ.CMX
24GCG.CMX
24GCJ.CMX
24GCM.CMX
24GCZ.CMX
25GCM.CMX
25GCZ.CMX
26GCZ.CMX
2021-01
2021-04
2021-07
2021-10
2022-01
2022-04
2022-07
2022-10
Date


Futures and Spot Volumes
Volume
23GCF.CMX
23GCG.CMX
23GCJ.CMX
23GCM.CMX
23GCQ.CMX
23GCV.CMX
23GCZ.CMX
24GCG.CMX
24GCJ.CMX
24GCM.CMX
24GCZ.CMX
25GCM.CMX
25GCZ.CMX
26GCZ.CMX
200000
175000
150000
125000
100000
75000
50000
25000
0
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
2022-07
2022-08
2022-09
2022-10
2022-11
Date


23GCF.CMX
23GCG.CMX
23GCJ.CMX
23GCM.CMX
23GCQ.CMX
23GCV.CMX
23GCZ.CMX
24GCG.CMX
24GCJ.CMX
24GCM.CMX
24GCZ.CMX
25GCM.CMX
25GCZ.CMX
26GCZ.CMX
2022-01
2022-02
2022-03
2022-04
2022-05
2022-06
2022-07
2022-08
2022-09
2022-10
2022-11
Date

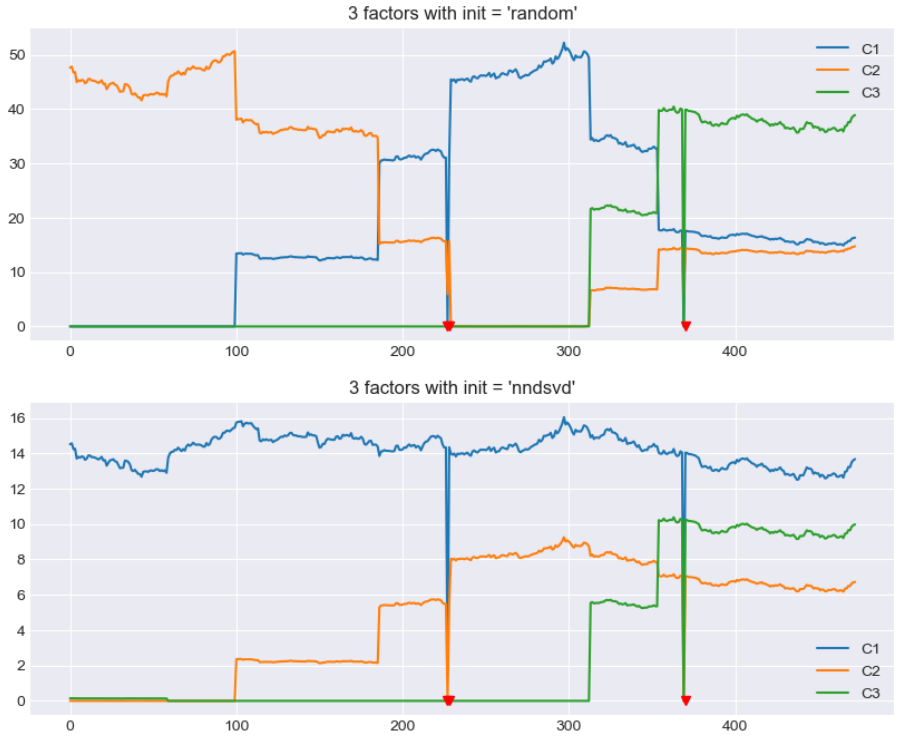
Price Analysis

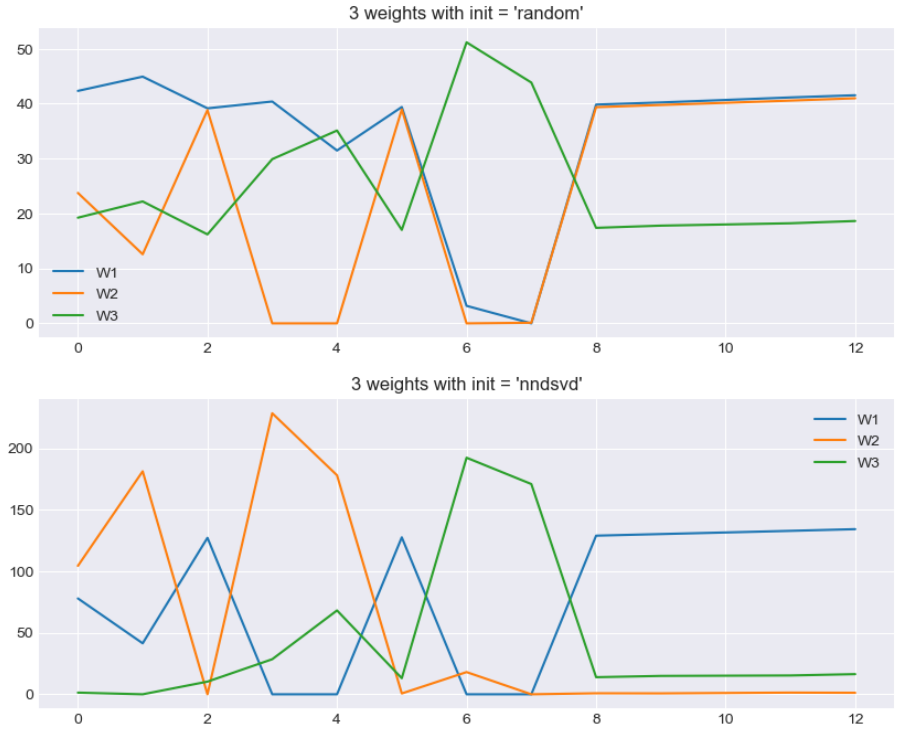
After conducting data pre-processing steps, we have the price data in a proper format; each column is the adjusted close price of a futures contract with different maturity. We can see that the instrument 'GCF23.CMX' has significantly fewer rows; thus, we will drop the data in the analysis; however, the missing data has a minor impact on the algorithm.

The three main driving factors of commodities futures instruments are the spot price, storage cost, and risk-free interest rate. Our inference lies in the hypothesis that we can reconstruct the three dynamics by performing matrix factorization.

First, the NMF algorithm may be divided into two main stages. One is the initialization of W and H matrices, and two is the update stage. Let us consider two versions of the NMF method. We will consider three components with the same beta-loss set to Frobenius norm for each step of the algorithm and random-state=0, but we will account for two different initializations. The first uses normal Gaussian priors, and the second uses nndsvd (Nonnegative Double Singular Value Decomposition). With the nndsvd technique, one performs two SVD processes. Many numerical examples suggest that the method drastically reduces the approximation error of several NMF algorithms. As seen in the beginning, our initial matrix of futures instruments is somewhat sparse; thus, setting the initialization stage with nndsvd method will be more helpful.

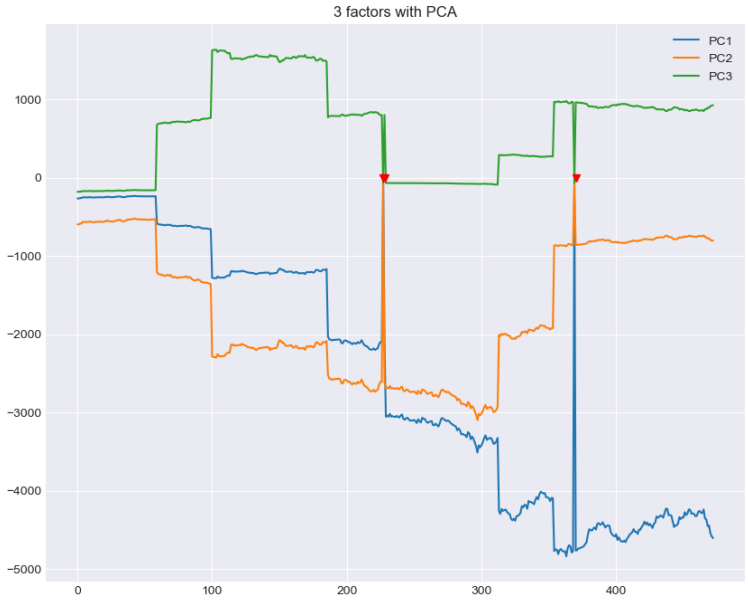
Filling the NA values with 0s, we obtain the following results for the factors in each case (Note: the sudden reduction of the components to 0 at a point (marked with the red triangles) is driven by the fact of filling the NAs with 0s):





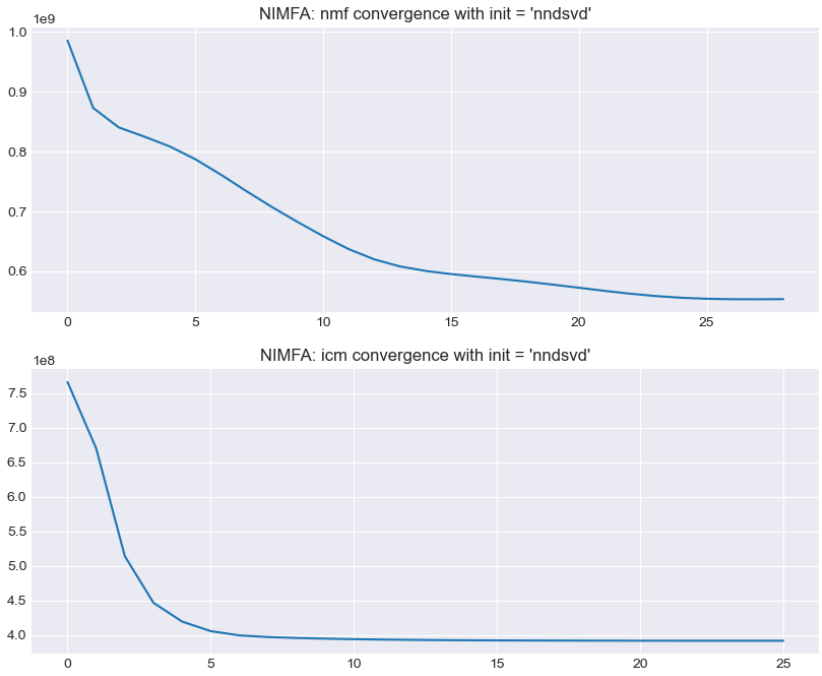
For this time, we may see a massive difference in the factors depending on the initialization method. One may notice that in the case of init='nndsvd', the factors seem more independent (if we say so). However, observing the feature weights plots, we may see some similarities in the patterns. For example, the 1st weight marked as W2 in the case of init='random' resembles W1 in the case of init='nndsvd', and the pattern of W3 resembles in both instances. In the 'random' case, we reach 145 iterations with a reconstruction error of over 19675, and in the case of 'nndsvd' we reach 200 with an error of over 19674.

From the PCA results seen below, the advantage of using the NMF method becomes obvious after constructing the primary components. As we are dealing with price data that is non-negative, the usage of NMF seems natural. However, compared to NMF components, the PCs are less intuitively interpretable, and even two of the PCs attain negative values. Notably, the spikes toward 0 occur in this case as well.

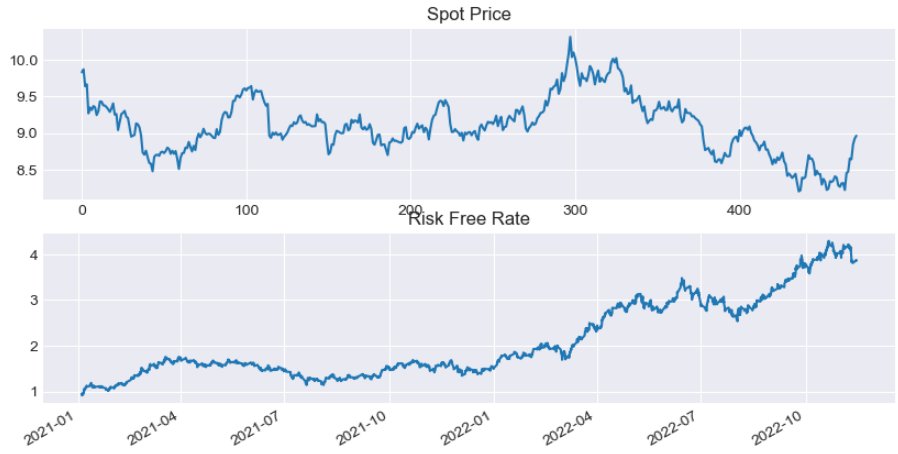


Now let us capture the components by utilizing another library called NIMFA; we will also account for the error at each execution step. In this case, we will run NMF and Iterated Conditional Modes non-negative matrix factorization (ICM) algorithms with the same settings. We will set the seed to 'nnsvd' as it performed better in the case of sklearn and utilize the update of divergence instead of euclidian and a metric of Frobenius norm. The graphs below show that the 1st component is the same as in the sklearn's case with the 'nndsvd' initialization step. However, it may not seem very easy to distinguish the other two components in the case of NMF of NIMFA from that of sklearn, but it is not the case of IMC. The IMC results replicate that of sklearn's with 'nndsvd.'

Also, comparing the convergence rates in both cases, it is seen that the convergence rate is faster in the IMC case. Moreover, the number of iterations of NIMFA is far fewer than that of sklearn, accounting for 29 and 26 for NMF and IMC cases respectively.

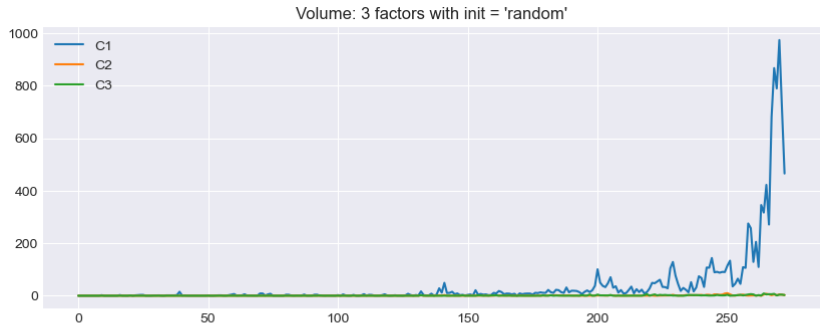


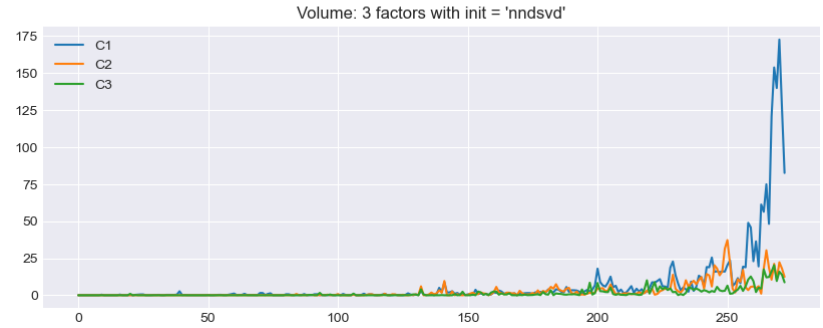
Finally, let us identify the components derived above. Bellow, we can see the spot price (GC=F) and the risk-free rate (the Treasury Yield 10 Years index) plotted. Observing both graphs closely, we find that the spot price is depicted as the 1st component in the cases of sklearn's with 'nndsvd' and NIMFA's. The components manage to capture significant bumps and lows precisely. On the other hand, the risk-free rate is depicted as the 3rd component for the sklearn's and NIMFA's cases by capturing the two sudden increasing trends seen at the end of the original plot. So, the leftover component showcases the storage cost dynamics. As the storage cost historical data was unavailable to acquire, we may claim it with the elimination method. Lastly, the PCA results cannot be interpreted as intuitively as NMF's.

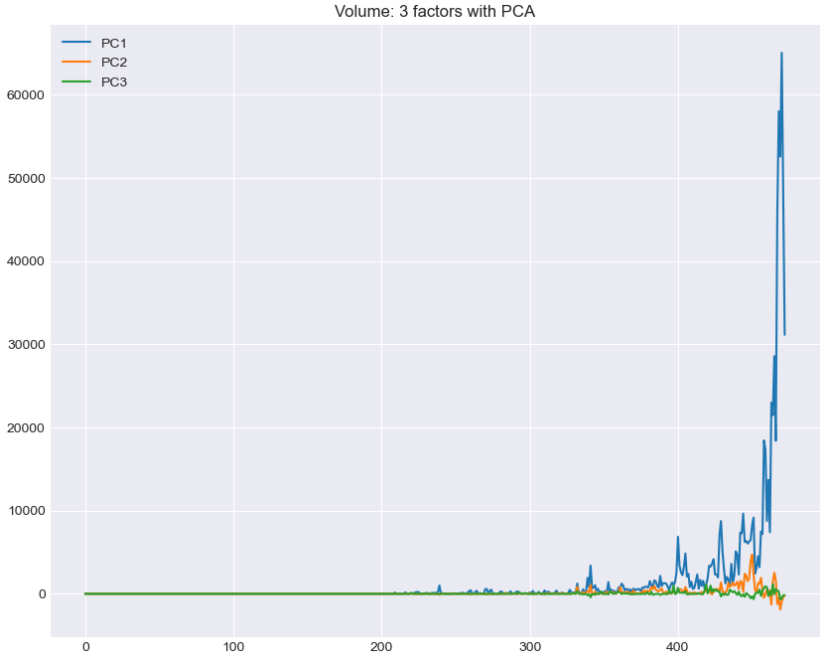


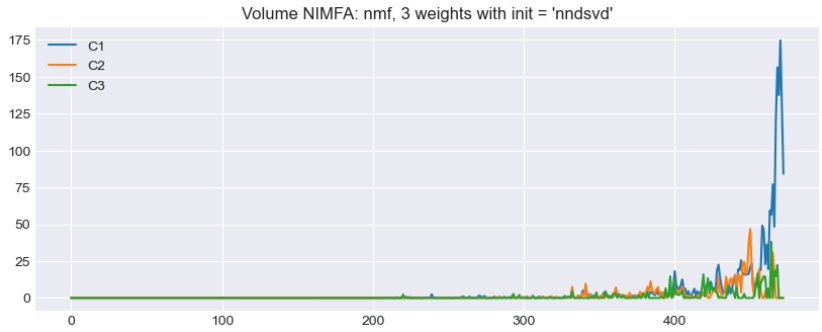
Volume Analysis

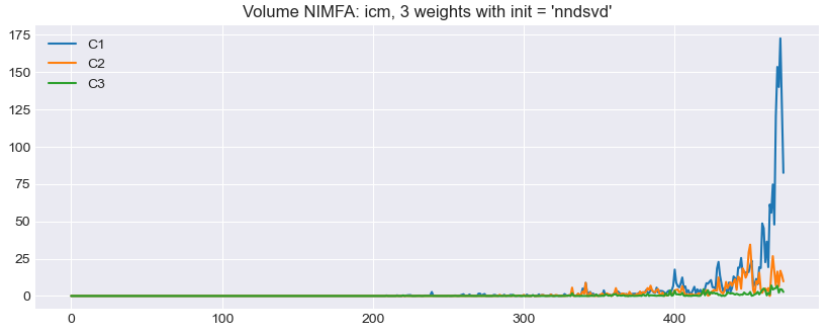
The pattern of the traded volume data seems obvious, as depicted below by the methodologies used. The traded volume pattern is steady throughout the lifetime of a commodities future but activates towards the end of it before the maturity date. The convergence rates can be found in the notebook and below in the graphs. Notably, the results obtained by NMF and PCA are comparable.

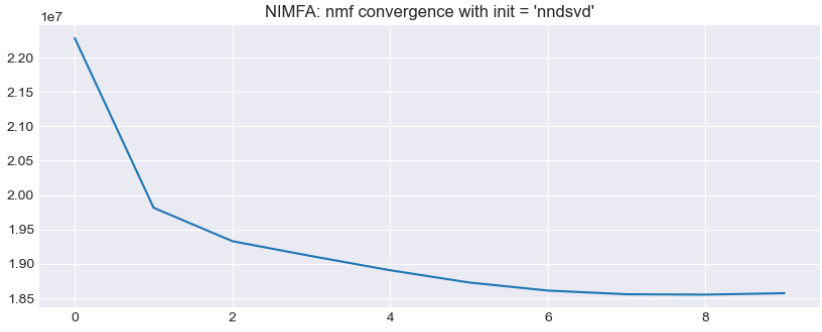


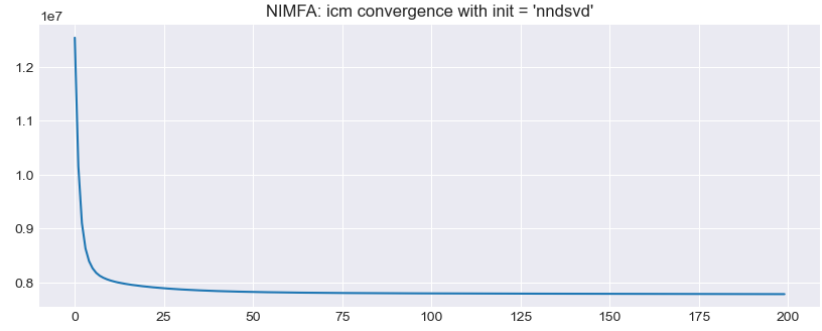





Volume NIMFA: nmf, 3 weights with init = 'nndsvd'
C1
C2
C3
175
150
125
100
75
50
25
0
0
100
200
300
400


Volume NIMFA: icm, 3 weights with init = 'nndsvd'
C1
C2
C3
175
150
125
100
75
50
25
0
0
100
200
300
400


NIMFA: nmf convergence with init = 'nndsvd'
1e7
2.20
2.15
2.10
2.05
2.00
1.95
1.90
1.85
0
2
4
6
8


NIMFA: icm convergence with init = 'nndsvd'
1e7
1.2
1.1
1.0
0.9
0.8
0
25
50
75
100
125
150
175
200