The success of many trading algorithms depends on the quality of the predictions of stock price movements. Predictions of the price of a single stock are generally less accurate than predictions of a portfolio of stocks. A classical strategy which makes the most of the predictability of the joint, rather than the individual, behavior of two assets is **Pairs Trading**, a Statistical Arbitrage strategy where a portfolio consisting of a linear combination of two assets is traded.

We will focus on the interaction between a tech stock (Intel, INTC) and a tech ETF (Merrill Lynch Semiconductor ETF, SMH) on November 5th, 2014. These two assets move together for two main reasons. The first is mechanical: around 20% of the ETF holdings are shares of INTC. The second is economic: the ETF is designed to represent the semiconductor industry, and hence its price will move in response to news that affects that industry, and the same news will have a similar effect on the price of INTC.

We assume that both INTC $S_t^{(1)}$ and SMH $S_t^{(2)}$ are assets whose dynamics have a transitory (mean-reverting) component and a permanent (Brownian) component. We express the dynamics of this mean-reverting (Ornstein–Uhlenbeck) process in vector form as follows:

$d\mathbf{S}_t = \boldsymbol{\kappa}(\boldsymbol{\theta} - \mathbf{S}_t)dt + \boldsymbol{\Sigma} d\mathbf{W}_t$ where $\mathbf{S}_t = \left[S_t^{(1)} \quad S_t^{(2)} \right]^\top \in \mathbb{R}^2, \boldsymbol{\kappa} \in \mathbb{R}^{2\times 2}, \boldsymbol{\theta} \in \mathbb{R}^2, \boldsymbol{\Sigma} = \boldsymbol{\sigma}\boldsymbol{\sigma}^\top \in \mathbb{R}^{2\times 2}$ and $\mathbf{W}_t \in \mathbb{R}^2$ is a 2d Brownian motion

The presence of a mean-reverting component introduces the opportunity for generating positive expected returns from trad ing by exploiting that component's predictability. In this case, we use the joint information from the two processes to create a stronger trading signal by constructing a linear combination of the two assets, which is most strongly driven by the mean-reverting component. This is done by transforming the system into an equivalent system:

$d\tilde{\mathbf{S}}_t = \tilde{\boldsymbol{\kappa}}(\tilde{\boldsymbol{\theta}} - \tilde{\mathbf{S}}_t)dt + \tilde{\boldsymbol{\Sigma}} d\tilde{\mathbf{W}}_t$ where $\boldsymbol{\kappa} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$ and $\tilde{\boldsymbol{\kappa}} = \boldsymbol{\Lambda} = \text{diag}(\lambda_1, \lambda_2), \tilde{\mathbf{S}}_t = \mathbf{U}^{-1}\mathbf{S}_t, \tilde{\boldsymbol{\theta}} = \mathbf{U}^{-1}\boldsymbol{\theta}, \tilde{\boldsymbol{\Sigma}} = \mathbf{U}^{-1}\boldsymbol{\Sigma}$

That is, apply eigenvalue decomposition on matrix $\boldsymbol{\kappa}$ and multiply $\mathbf{U}^{-1}$ on both sides of the stochastic differential equation.

Don't worry if you find the above model difficult to understand, it is just for your reference.

Our goal is to find a linear combination of $S_t^{(1)}$ and $S_t^{(2)}$, expressed as $\tilde{S}_t = c_1 S_t^{(1)} + c_2 S_t^{(2)}$ such that it will have the strongest exposure to the mean-reverting behavior of the two assets and it happens that such transformation is under the eigenbasis of $\boldsymbol{\kappa}$ (analogous to Principle Component Analysis, aka PCA).

Step1. Estimate $\boldsymbol{\kappa}$ from the data.

To estimate the discrete version of our model, we estimate the vector regressive (VAR) process: $\mathbf{S}_t = \mathbf{A} + \mathbf{B}\mathbf{S}_{t-1} + \boldsymbol{\varepsilon}_t$ where $\mathbf{A} \in \mathbb{R}^2, \mathbf{B} \in \mathbb{R}^{2\times 2}$

It is equivalent to estimating two linear regression models: $S_t^{(1)} = a_1 + b_{11}S_{t-1}^{(1)} + b_{12}S_{t-1}^{(2)} + \varepsilon_t$ and $S_t^{(1)} = a_2 + b_{21}S_{t-1}^{(1)} + b_{22}S_{t-1}^{(2)} + \varepsilon_t$

Recover the parameter $\boldsymbol{\kappa}$ of our model by $\boldsymbol{\kappa} = (\mathbf{I} - \mathbf{B})/\Delta t = \mathbf{I} - \mathbf{B}$ where Δt is set to be 1 minute in our data.

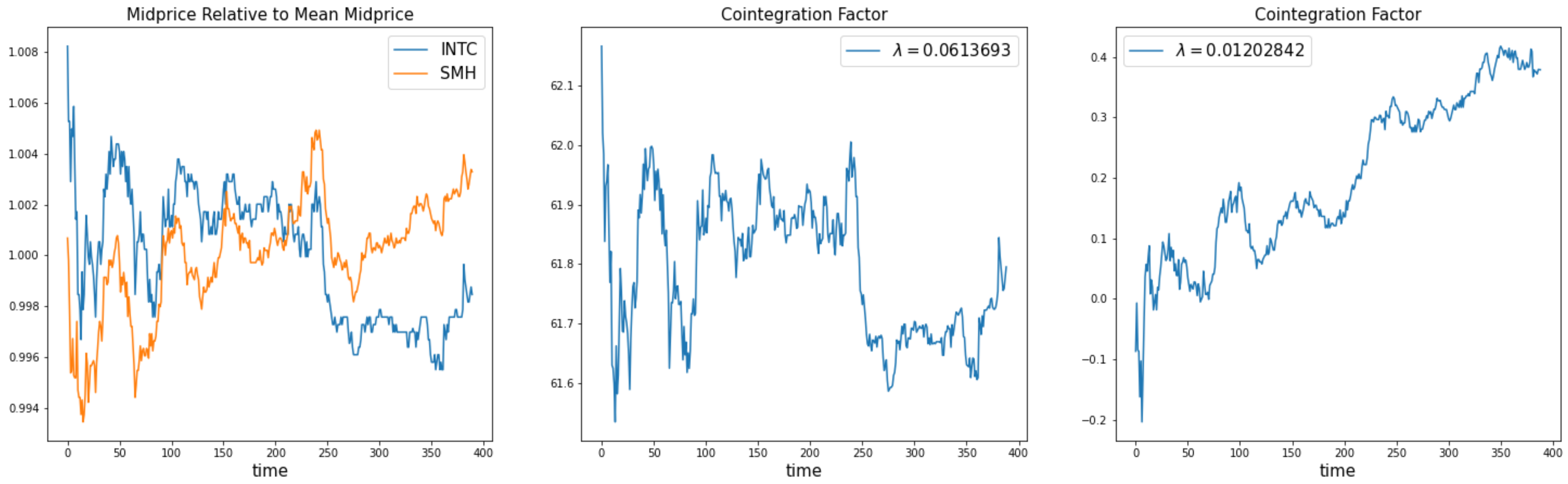
Step2. Perform eigenvalue decomposition on $\boldsymbol{\kappa}$ and obtain transformation matrix $\mathbf{U}^{-1}$

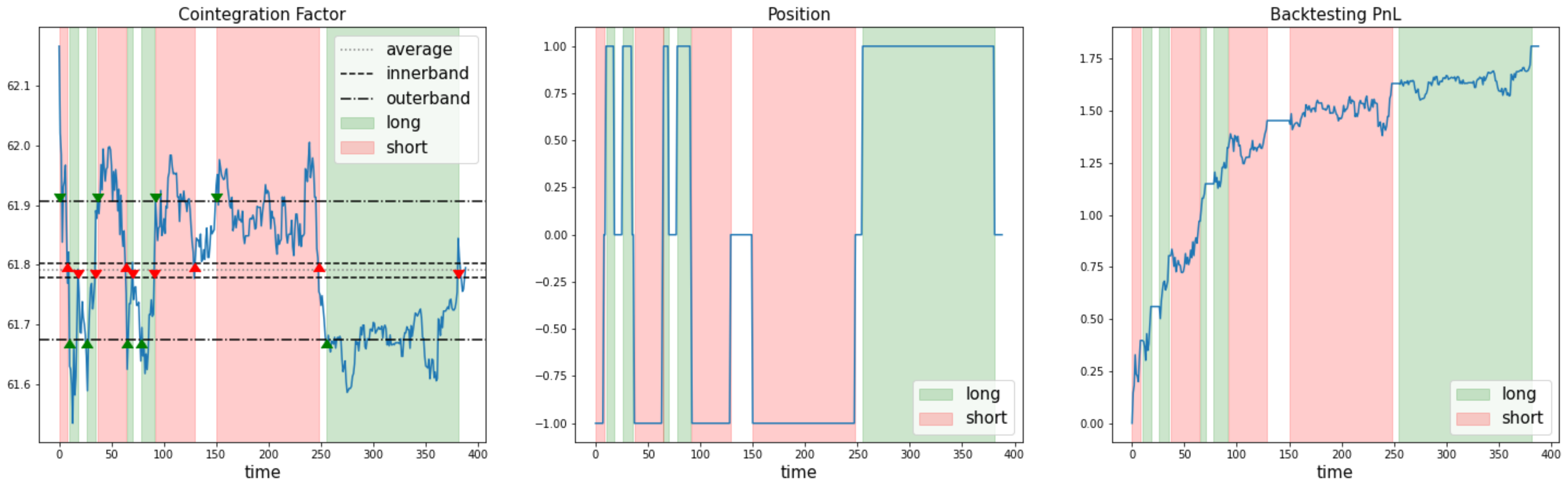
Decompose $\boldsymbol{\kappa} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{-1}$ and pick the process $\tilde{S}_t$ from the vector process $\tilde{\mathbf{S}}_t = \mathbf{U}^{-1}\mathbf{S}_t$ that is corresponding to the largest eigenvalue.

That is, $\tilde{S}_t = \tilde{S}_t^{(j)}$ where $j = \arg\max \lambda_i$.

Step3. Backtest.

Assume that from above we get $\tilde{S}_t = c_1 S_t^{(1)} + c_2 S_t^{(2)}$, construct 1 contract of co-integration portfolio $\tilde{S}_t$ by long/short c_1 shares of $S_t^{(1)}$ and long/short c_2 shares of $S_t^{(2)}$.
Based on our mean-reverting assumption, we develop a simple statistical arbitrage strategy that short this portfolio when its value is too high and long it when its value is too low.
Introduce two bands as reference line for the value of this co-integration portfolio, the inner bands is $\text{avg}(\tilde{S}_t) \pm m$ and the outer band is $\text{avg}(\tilde{S}_t) \pm M$ with $m < M$.
When the value reaches outside of the outer band, we will open a long or short position and we will close the position when the value reverts and hit the inner band, illustrated as follows:





Some Remarks: This is a very simple but naive strategy for entering and exiting a long or short position in the co-integration portfolio. There is also an information leakage problem because the inner and outer bands are calculated with all the information of prices of the day we backtested on. In production more sophisticated control theories will come into play for setting the optimal bands and if you are interested, please refer to Chapter 11 of Book Algorithmic and High-Frequency Trading for more information. In this problem, however, for simplicity we just assume the bands are already given.