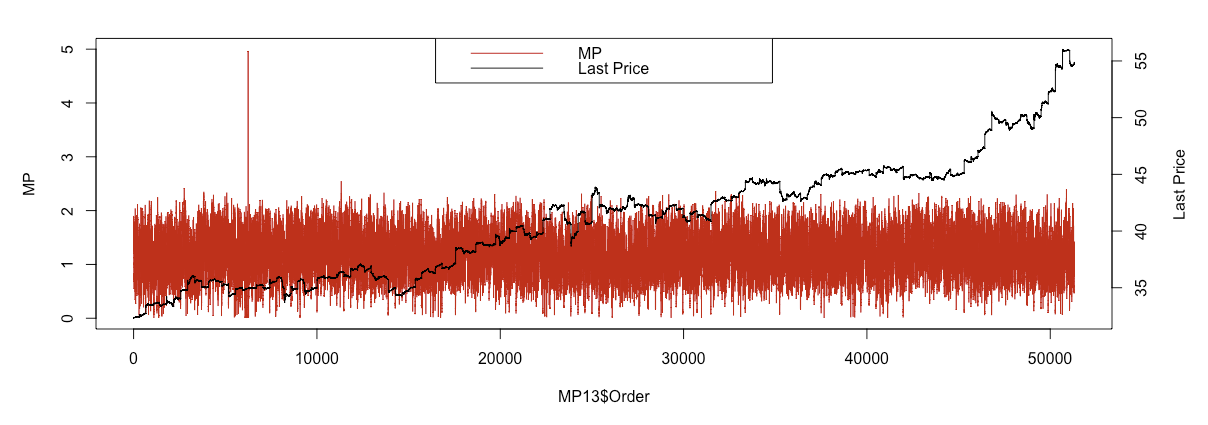
Matrix Profile serves as a unique way for us to exam the data and to facilitate stock investment strategy. It’s a breakthrough for all pairs similarity joins for time series. The only parameter in this method is the size of the time series subsequence, m, which is defined by users. The original time series, which has n observations, is reconstructed into a new sequence with n-m+1 as the number of subsequences. The first subsequence contains the observations from 1 to m in the original time series and the second subsequence includes the observations from 2 to m+1, and so on. Then we calculate the z-normalized Euclidean distance of each subsequences with all the other subsequences, and return the smallest distances as matrix profile and location of its nearest neighbor as matrix profile index.

The method is focused on local properties of time series. It helps many time series data mining problems, including motif discovery, discord discovery, shapelet discovery, semantic segmentation, density estimation, and contrast set mining. One of the major contributions of Matrix Profile is its great improvement of calculation speed so that the join essentially could be maintained on a standard desktop even if the data arrival frequency was much faster than once a minute. This feature will give us the ability to provide timely estimation on minute trading activities of stocks.

There are also other advantages of this method:

* It is exact, providing no false positives or dismissals.
* It is parameter-free so that there is no over-fit problem.
* It is extremely scalable and even for extremely large datasets, the results can be provided in a anytime fashion with ultra-fast approximation.
* The method provides full joins, eliminating the need to specify a threshold.
* The algorithm is parallelizable, both on multicore processors and in distributed systems.

We calculated the matrix profiles form from 4 to 60 on LAST PRICE of Tencent stock, and we use subsequence size 13 as an example to verify and demonstrate the effectiveness of Matrix Profile on the dataset we studied. The graph below is a plot of Tencent’s Last Price and matrix profile outputs from May 2017 to November 2017.



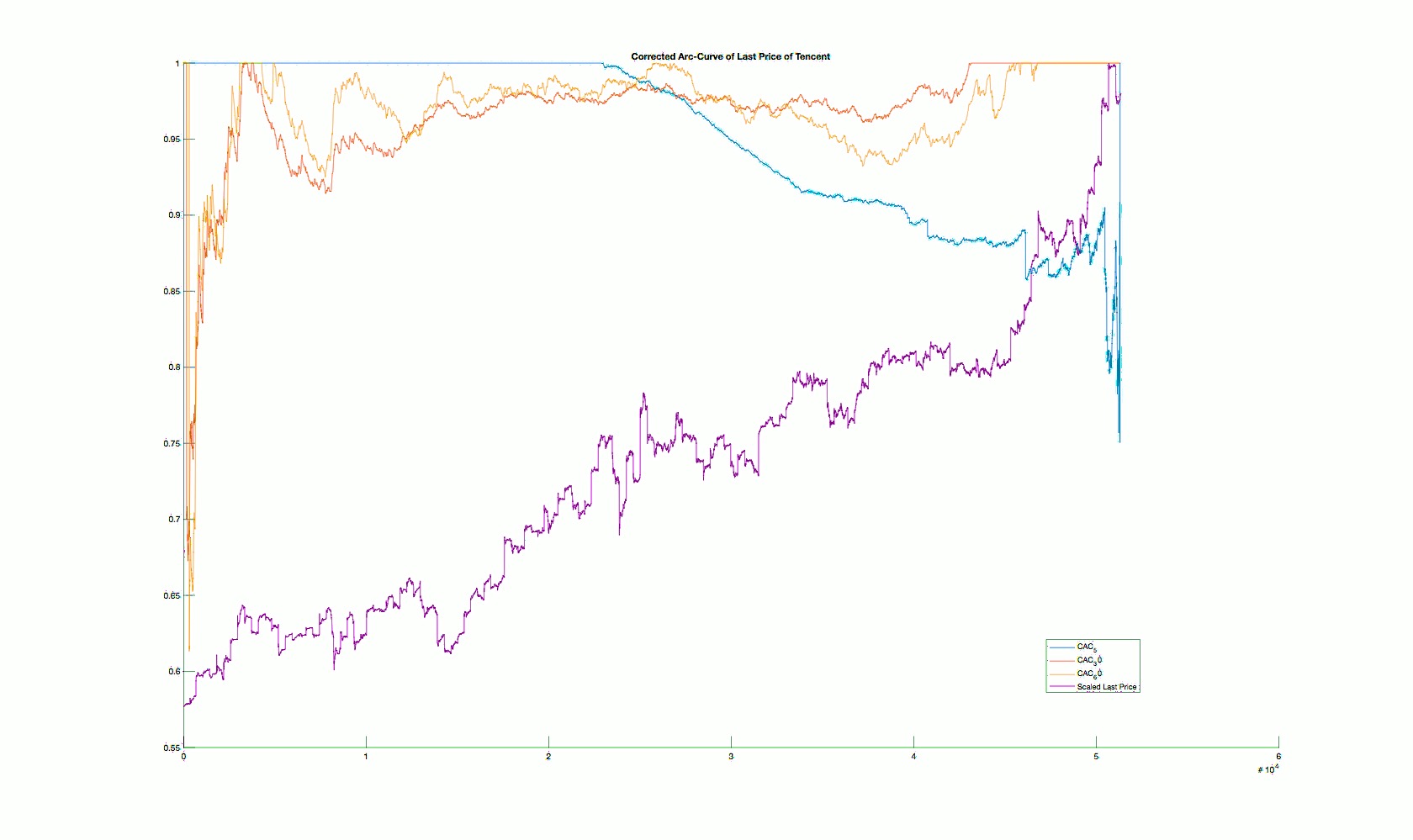
Since matrix profile is the distance of each subsequence from its nearest neighbor, it’s intuitive that the spike shows the existence of discord and the near zero point means the existence of motif. From the graph above, we can easily find the highest spike and some lowest points.

We first exam its discord discovery feature by plotting the highest spike against matrix profile (see graph below). The discord happens between 6/2/2016 14:34 to 6/2/2017 14:47 during which period Tencent’s last price stays the same. There is no material news during this time period, which will drive a significant price movement. After taking a close look at the number of ticks, which means the number of trades in one minute and trading volume, we further confirm the rationality of the discord. There are only 2 or 3 ticks in each minute over this period, which is lower than the mean of the number of ticks over the whole period studied and it could be possible the trades are from the same buyer, whose order gets split due to lack of supply. 

We then output the smallest point of matrix profile and its corresponding neighbor in the graph below. The pairs are last prices from 6/2/2017 11:58 to 6/2/2017 12:11 and 8/18/2017 14:50 to 8/18/2017 15:03. It is clear to us that the behaviors of the two time periods have a very high similarity. The rationality of the best motif is verified.



The corrected arc-curve is a alternation of Matrix Profile outputs and results in a discovery of shift of regime. It is a parameter-free method. Arc-curve counts the number of nearest neighbor pairs after each data point. The beginning and ending of a time series will have less number of neighbors so the corrected arc-curve(CAC) does corrections and scaling on both sides. It means the existence different segmentations when the output is near zero. The method is not sensitive to the subsequence size, m. We plot the CACs with m equals to 5(in blue), 30(in orange) and 60(in yellow), and the scaled LAST PRICE of Tencent(in purple). There is no doubt that the whole time series has no regime shifts since all the CACs stay fairly away from zero. The finding matches our independent study of the Tencent stock price behavior using Bloomberg. This gives us the confidence in modeling the entire time series as a whole.



The histogram of the matrix profile also provides an answer to the time series density estimation. The variance of the matrix profile is a measure of the time series’ s complexity.

It is convincing that matrix profile contains useful information of the data. A plot of matrix profiles with different m shows different structure (see graph) and it provides different views of the dataset. So it is important to take into consideration matrix profiles with different subsequence sizes.



Source: Matrix Profile I: All Pairs Similarity Joins for Time Series: A Unifying View that Includes Motifs, Discords and Shapelets. Chin-Chia Michael Yeh et al (2016). IEEE ICDM 2016.