Description of Assignment A

Yixiang Wang

The assignment can be divided into three parts:

1) Parallel data I/O with MPI
2) Scrub the original data set to separate signal and noise
3) Test the statistical characteristic (mainly normality) of the signal data.

This brief introduction file will detail each of the three parts respectively.

PART ONE: Parallel data I/O with MPI

Though parallel computing can be implemented on different levels, for assignment A, only MPI level parallelism is used. And data I/O is actually the only part that we need to use the mpi library (mpi.h).

Data input in this assignment is relevantly easy. The program scrub will take only one input file. When calling the scrub program, the user need to specify the number of nodes available. The main function will call MPI_File_get_size() to get the size of input, equally distribute the data to all nodes, and calculate the starting points and offsets of each node. Then MPI_File_read_at() function is used to read the data none collectively since the scrubbing jobs of each node are basically independent.

Note that some records at the boundary of two parts corresponding to different nodes may be incomplete. For simplicity, the program will just throw it away since one single record should not have large impact on the analysis.

The file output is a bit trickier than the input, since in this case each node needs to know the output size of other nodes to determine its own writing starting point. This job is done with MPI_Allgather() function.

The program will output five files: signal.txt, noise.txt, performance.txt, normality test result.txt, as well as a log file recording the program activities.

PART TWO: scrubbing data

Data scrubbing is mainly completed with sliding window methods. After read the data from input file, each node will call string2record() function to convert the input string to a vector of record, which is defined as a struct with three elements: time (which is also a struct), price and size.

To begin scrubbing, a sliding window with length of 500 will be initialized. The record in the sliding window will be sorted at the beginning, which will be fast given the small size of the window.

The scrubbing will take two pass of the vector, each serves different purpose. The first pass will use a sliding window to put records in locally correct order, separate the records that is out of order too far away (earlier than the last record in the sliding window), and also calculate the mean and standard deviation of signal. The second pass will use the standard deviation and mean calculated in the first pass to further select noises. Any record with a price deviated from mean by greater than three standard deviations or with a negative size will be considered as noises.

PART THREE: normality test

Many algorithms are available to test the normality of a return sequence, which is defined as log return in this assignment. Here I choose Jarque-Bera test since it does not require to calculate the sample quantile and can thus avoid sorting the whole array. The JB statistic can be written as:

$$JB = n \times \{\frac{\widehat{sk}^2}{6} + \frac{(\widehat{kur}-3)^2}{24}\}$$

Where $\widehat{sk}$ and $\widehat{kur}$ are sample skewness and kurtosis. For large sample size, JB statistic has a χ^2 distribution of order 2. Since even the smallest sample tested has thousands of records in our case, we can assume the JB statistic in our case are χ^2 distributed and thus can be tested with one critical value. We choose confidence level to be 0.95 and reject null when JB statistic is greater than 5.991465.