MTH 9898 Big Data in Finance

Assignment A

Ruida Xie

Building a Real-Time Data Scrubber and Analyzer

Overview

Using parallel computing, the program reads data from .txt file at multiple starting points. It conducts the required data processing, including flagging problematic data and test normality of log-returns, and outputs valid data and noise data. In command line, under the file directory, run “mpicxx Assignment\_A.cpp –std=c++11” and the program will work just fine. Around 10% to 15% of the data are noise. To test the normality of the given data, the program uses Jarque-Bera Test. None of the test data is significantly normal under this test.

Data Scrubbing

The program reads data line by line. In every line, test if there are symbols other than digits, colons, commas, and dots in the text. Test if the time is well formatted and if any one time is close to its neighboring times. Test if the volumes and the prices are far larger or smaller than their neighbors. Output data with these problems to noise\_data.txt and the rest to valid\_data.txt. The program marks five timestamps: the start of the program, the start and the end of the reading progress, and the start and the end of the writing progress. Therefore, the users can track the progress and infer the duration of both the reading and writing part.

Note: If there is an error while loading shared libraries, running the following command lines will fix the error.

$ echo $LD\_LIBRARY\_PATH

$ LD\_LIBRARY\_PATH=/usr/local/lib

$ LD\_LIBRARY\_PATH=$LD\_LIBRARY\_PATH:/my\_library/filename.so.version

$ export LD\_LIBRARY\_PATH

Normality Test

Because the sample size of the input data is supposed to be extremely large, the skewness and the kurtosis of the log-returns should be very close to 0 and 3 respectively if they are normally distributed. The program will output the skewness and the kurtosis in the console window, and none of the values match the requirement of normality.