Phase 2 Program Documentation

Location of program

`/space/mth693/common/phase2_programs`

Notes

All programs are designed to keep asking for input until a valid one is inputted. To end the program at anytime, you can press CTRL+C to raise KeyboardInterrupt and stop the program.

Before Running the Program

Type the following in the terminal to install the R packages needed for the ARIMA based method:

`Rscript setup.R`

Running the Program

Once you are in the phase 2 program directory, before running the program, type the following command into the terminal:

`source /space/mth693/common/phase1_programs/crypto/bin/activate`

You should now see (crypto) to the left of your name. The above command starts up a python virtual environment. The virtualenv has the required packages to run the programs. To run a python script called `myprogram.py`, type the following:

`python myprogram.py`

Program Description

All output will go to the user's workspace folder, i.e. ~/workspace.

phase2

Filename: `phase2.py`

Description:
This program gets the statistics (num outliers, percent outliers, etc), outliers and its "strength" as a csv, and a plot of where are the outliers.

Notes about strength:
For k-gamma: the larger the number, the more likely it is an outlier
For isolation forest: the smaller the number, the more likely it is an outlier
For ARIMA based, the larger the number (t-statistic), the more likely it is an outlier (loosely speaking)

Inputs:
`filetype`: A (1619 ICE Data Feed) or B (1356 Blockstream)
`date`: the date in yyyymmdd format
`ticker`: the pair to plot
`option1`: Trade, Bid, or Ask
`Venue`: the venue
`Method`: 1, 2, or 3 for k-gamma, isolation forest, and ARIMA based respectively

Parameters: depending on the method chosen
K-gamma
      k (int): window size
      gamma (float): granularity parameter
Isolation forest
      contamination (float): fraction of the total data with outliers, as a decimal (eg 0.1 for 10% outliers)
ARIMA based:
      critical value (float): the threshold for the t-statistic. If a suspected point has a t-statistic lower than the critical value, then it is not an outlier

Output: For all outlier detection methods. A .txt file with the statistics of the method and the outliers. A .csv file with all of the detected outliers and their "strength". A .png file with the location of outliers.

After running program

Type deactivate to deactivate the virtual environment.

Advanced Usage:

The program is designed to take in command line inputs, so that you can input the parameters in the same line as the python command (skipping the prompts up to the method parameters), rather than inputting the parameters after you start the program. For example, you can type:

```
python phase2.py A 20180628 X:SXBTUSD Trade CTC 2
```

rather than type `python phase2.py`, then input those parameters in by hand (note that you do not input the outlier detection parameters). If you have the parameters saved in a file called input.txt, then you can type:

```
python phase2.py $(cat input.txt)
```