Statistical Arbitrage MGMTMFE 431/3 – Professor Olivier Ledoit – UCLA Anderson School

Problem Set #1: Universe

WARNING: This problem set is very long. You have one week to do it. It is due on Thursday October 10th, 2019 before class (08:30am atomic clock). If you start two days before you'll never make it. Start today.

The point of this exercise is for you to be resourceful in finding the data you need. You need to be able to "manage the complexity" of acquiring and organizing so much data. This is much of the difficulty of Statistical Arbitrage strategies, and a necessary skill if you want to succeed.

1) We will work with the following list of 13 European stock market indices:

- Belgium: BEL-20
- Denmark: OMX Copenhagen 20
- Finland: HEX-25
- France: SBF-120
- Germany: Xetra DAX-100
- Holland: AEX-25 and AMX-25
- Italy: MIB-30 and MIDEX
- Norway: OBX-25
- Spain: IBEX-35
- Sweden: OMX Stockholm 30
- Switzerland: SMI-20

For each one of these 13 stock market indices, retrieve the list of the DataSteam codes (dscodes) of its constituents on each of these 5 dates:

- December 31st, 1997
- December 31st, 1998
- December 31st, 1999
- December 31st, 2000
- December 31st, 2001

If historical constituents for a given index are unavailable prior to, say, July 31st 1999, then you are allowed to backfill the December 31st 1997 and December 31st 1998 data with the July 31st 1999 list. We will ignore the forward-looking bias that it induces. But you must go as far back into the past as you possibly can.

HINT:

- We recommend that you install the DataStream Excel add-in on your account when you use DataStream in the library lab – rather than use the stand-alone DataStream front-end.

2) Form a merged list of all unique dscodes. Let `n` denote the length of the merged list.

3) Create a Matlab structured array of dimension `(1×n)` called `allstocks` such that `allstocks(i).dscode` is a string equal to the dscode of the i^{th} stock on the merged list (for `i=1,...,n`).

4) Within the variable `allstocks`, create a sub-structure called `namelist`:
`allstocks(i).namelist`

Set `allstocks(i).namelist(1).date = '01-Jan-1998'`
Set `allstocks(i).namelist(1).name` equal to the current name of the `i`th stock.

If the name of the `i`th stock changed, say, on November 8th, 2016, then we would set `allstocks(i).namelist(2).date = '08-Nov-2016'` and set `allstocks(i).namelist(2).name` equal to the new name of the `i`th stock as from November 8th, 2016. Thus, this structure is very flexible. That is what you would do in the real world. But, in the interest of making the assignment less difficult, we will just have one name per stock.

5) In a similar way, populate `allstocks(i).industrylist(1)` using the 5-letters DataStream level-4 industry mnemonic.

6) In a similar way, populate `allstocks(i).ibeslist(1)` using the I/B/E/S ticker. There should be only one I/B/E/S ticker (which is *not* the official exchange ticker!) per dscode, but we have to be ready in case this rule is violated.

7) In a similar way, populate `allstocks(i).indexlist(1)` using stock index memberships: for example `'XDAX100'`, or `'FBSF120'`, etc. For the two Italian indices, use `'ITMILAN'` and for the two Dutch indices use `'AMSINDX'`. If a stock does not belong to any index at the beginning of year `j`, then set `allstocks(i).indexlist(j).index='';` (empty string).

8) In a similar way, populate `allstocks(i).isinlist(1)` using 12-character alphanumeric ISIN codes.

Note: For items 5) to 8), the expression "In a similar way" implies, in particular, that you need to have a sub-structure with the dates: it should be organized like item 4), except covering different information.

Problem Set output: Your group must submit in CCLE a Matlab database called allstocks.mat containing the variable `allstocks`. CCLE will lock-down this assignment at 08:30am on Thursday October 10th and no further submissions will be possible.

Grading: You and your group members will be judged by how complete and correct `allstocks` is.

I am nice: In order to simplify the assignment, you are allowed to retrieve only one name, one industry mnemonic, and one ISIN code per stock. Just take the current ones. In theory, these things could change over time, and we should keep track of those changes, but I authorize you to just get one per stock. Note that you must construct the list structures exactly as stated above: it allows for changes in all the fields over time, even if there is only one item in each list currently. Only the index membership list can have multiple items as stocks go in and out of the index constituent lists.