Hypothesis 1: Every index on a stock exchange can be presented as a combination of several other indexes (no time-dependency considered)

December 4, 2014

1 General Idea

Here is general idea broken into several steps is described.

1. Main Index is chosen ->X
2. The number of indexes will be limited through: choosing indexes from related markets, indexes from the same market. (no unexpected cross-correlation)
3. The number of possible indexes is almost not limited - only impossible or weakly correlated markets are excluded. Chosen indexes and corresponding time-series are saved to an array for further use.
4. For the main index X "best-match"-regression will be found through iteration over a number of correlated indexes. The model will be manually limited to N indexes that result in closest trend in comparison to the original index X.
5. For these indexes
$$x_0, ..., x_n$$
texts will be found and analysed, as the output we will get N lists (tuples)
$$[probability,\ forecast]_i,\ where\, i = \overline{0, n}$$
6. Afterwards all these forecasts will be integrated in one general forecast for main Index X

This hypothesis mostly aims already existing companies, listed on stock exchanges.

2 Trivial and non-trivial sub-tasks

Probable problems, bottle-neck places and trivial tasks of the work.

2.1 Trivial tasks

Trivial sub-tasks are steps that are easy to implement, they are normally transparent and run-time is not an import issue.

1. Crawlers don't have to contain time-parameter (it is somewhat hard to estimate the correct time-period in the past for future without explicitly given time-frames of the forecast).
2. The dependencies (at least dependencies with the biggest impact on each other) are mostly established and known. So, probably the amount of indecies that we need to check with decided regression-function will be relative small.
3. Also the sub-markets of the indexes are relatively easily separated. And the cross-market correlation in the same time-period should be insignificant.
4. Technically the integration of the separate

   $$forecast_i$$

   should be easier, because of no time-dependency (lags) in the model.
5. Based on the previous assumption an interpretation of the results should be easier (the error-finding should more transparent as well)
6. Mood and buzz do not play a crucial role here. Forecast is the needed output format.

2.2 Non-trivial tasks

Here most obvious problems of the implementation are announced.

1. The input data - where to take the information for model-choosing (indexes, time-series, forecasts)
2. There is no way to state that the list of the indexes we will iterate over, is likely to be complete and no significant information is lost.
3. How to choose "best-match"?
4. Iteration over big number of parameters with the big number of variables (the second number is given explicitly and manually in the beginning, based on "intuition").
5. The classical regression-model is most likely insensitive to market-shifts and economical environment changes (scaling problem).

3 Conclusion

1. Positive case
   This hypothesis is a general assumption, that can give a hint where to look, in order to get additional advantage. Potentially it can be implemented as a module for example for StockPulse, that will deepen the analysis.

2. Negative case
   In this work we will only consider specific sub-market. And the negative hypothesis can be result of the wrong regression-model, or low number of parameters and no cross-market influence included in the consideration (in this case the model should become larger and complexer).

Independent from the result, this work can be helpful in choosing the information for analysing the stock exchange.