Hypothesis 2: Every index on a stock exchange can be presented as a combination of several other indexes and some of these variables have a postponed impact (Regression models with lags).

December 18, 2014

1 General Idea

The Idea is to check whether indirect sentiment data analysis will result in a higher accuracy forecast or even time-advantage for traders. By indirect analysis we mean the analysis of the variables from the regression for the main index. For example we choose to get a forecast for index X, the "best match" regression with lags will be found:

$$X = F(w_{it}),$$

where

$$i \in I$$

, I- is a number of the variables included,

$$t \in T$$

, where T- is the time of the influence.
For chosen variables we will get a forecast and than based on these forecasts we will reconstruct the forecast for the original index X.
Below is the whole process described more detailed and broken into logical steps.

1. The main Index is chosen ->X
2. The main correlated markets are limited (here we will consider several dependency-indicators and choose those markets, whose indicators are below certain level)
3. The shares listed on the chosen markets and their time-series over certain period and certain frequency (here depending on the model we will either

take high frequency or low frequency data) will be stored 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+m indexes. The combination of these indexes should result in the closest trend in comparison to the original trend of the index X.

$$X = f(x_{i_{t_j}}),\ where\, i = \overline{0,n},\ and\, j = \overline{-T,0}$$
$$T, n \in N$$

here N indexes are manually given amount of variables for the regression, if the model does not hit desirable or acceptable error ratio, than the "tail" will be increased maximum for m (also limited above) variables. If the model doesn't hit acceptable ration even with N+m variables, another model will be considered. (certain number of regression-models or some other tools will be pre-decided based on researches)

5. For chosen sub-indexes:

$$x_{i_{t_j}}$$

forecasts with it's probability will be taken - here stock pulse program could be used.

6. Final forecasts for variables will reconstruct the final forecast for the original index X.

7. The results will be compared with actual value of the share on the market in T+1 step. (Test sample)

Our forecast will be compared with the forecast of the StockPulse. Here the first conclusions can be maid:

1. If the indirect forecast matches the direct forecast (or has an insignificant difference within certain error ratio), then within the given models and parameters restriction indirect analysis does not give any additional accuracy or time-advantage.

2. If the indirect forecast is better than the original forecast, than within given models and variables restrictions the approach results in additional advantage (for StockPulse it means higher accuracy predictions, which increases the value of the data and the platform in general)

The indirect forecast can't be worse than the original forecast, because it will be trained and tested on the real data. In case the idea doesn't work and results in a forecast with a huge error (not even comparing with the forecast of StockPulse), than the hypothesis for given restrictions will be rejected and no further coamparison will be done.

2 Trivial and non-trivial sub-tasks

The research itself aims the hypothesis. In case the hypothesis is positive in most cases (specify "most cases"), than the idea can be implemented as an automatic evaluation-test for indexes. This test can be implemented as a module for the StockPulse platform.
There are of course some difficulties with possible implementation. Below will be given the most obvious trivial and non-trivial tasks for simple case (hard restrictions).

2.1 Trivial tasks

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

1. The step with the text processing is excessive.
2. Here buzz and mood can be used as "non-explicit" indicators for fine-tuning.
3. Since we are not working with texts, but with forecasts directly we will most likely find
$$[probability,\ forecast]_{i_{t_j}}$$
without a problem.

2.2 Non-trivial tasks

Here most obvious problems of the implementation are announced.

1. The first problem is to choose number of the indexes we will later iterate over in order to find the closest "representatives". Here not only cross-market correlation should be considered, but also time-dependence on these correlations.
2. Storage shortage can occur, when we consider too many variables. But it is not an issue for already existing database of StockPulse (the need data is a sub-data can be easily extracted)
3. The task of finding the optimum is hard, because of time-dependencies and no obvious dependencies. Here proper objective function needs to be found and also some additional error-estimators.
4. Iterating through list of indexes should also contain restriction on cross-correlation among those indexes.
5. Integration and interpretation of the forecasts can be hard and not transparent.

6. Run-time and complexity, scaling - all these can be problematic. In the final version multy-threading implemented in c++ (currently best run-time) is the best variant (complexity of the implementation and probably some difficulties with compatibility).

3 Conclusion

If the hypothesis is positive in most cases, then the additional advantage can be observed not only for existing companies, but also for brand-new products launches and IPOs.
This hypothesis aims both existing companies listed on the stock exchange and brand-new companies, know-hows and IPOs - the cases, where people will most definitely (under-) overestimate the company, because of lack of the information. Since the task is to obtain hidden or indirect information this approach should be especially helpful for estimation of the IPOs, psychological factor tend to give significant boost to opening prices.