Prediction of the development of a stock index based on its most significant constituents

Report author: Miquel Marti i Rabadan *miquelmr@kth.se*
Group partner: Philippe Weitz

January 11, 2016

Abstract

A particle filter approach to the prediction of stock indexes is analysed and implemented with a measurement model based on a weighted average of the most significant constituents of the index and different process models. A process model mimicking a random walk which gives no improvement on the prediction using the partice filter over the measurement model itself and a process model in which a trend estimate is introduced, giving a better prediction of the stock index by the particle filter when the trend estimate is correct.

1 Introduction

Stock indexes represent not only the real situation of stock markets (or sectors) but can also be considered as bellwethers of the whole economy of a country. Prediction of the future behaviour of those could be useful not only for gaining profit with market trading but for planning politics, business operations and others.

Stock indexes are normally weighted averages of different stock prices which are said to be their constituents. The weights are computed in a number of ways, being the most common based on price (Dow Jones Industrial Average) or capitalization (Standard&Poor's 500 Index). How the constituents are selected also varies from index to index: for the DJIA the

selection is made by the editors of the Wall Street Journal on the 30 companies that best represent the U.S. stock market, for the S&P500 the 500 largest companies of the U.S. stock market are selected. Stock indexes cannot be traded on, but exchange-traded funds or ETFs normally try to track those by aggregating stock shares in a similar way than is done by the indexes.

A particle filter approach to the prediction of the S&P500 stock index is analysed, based on estimates about the most significant constituents of the index, which are assumed to be very precise. This could be the case if the investor had insight information of the company and very good prediction schemes for these reduced number of stocks.

Publicly available stock data of the index and its constituents is analysed in order to find the most relevant symbols, which are then used to train a measurement model for the index based on a multiple linear regression. Different process models are proposed and tested based: a first model just adds noise to the previous particles while the second includes a prediction on the trend for the following days. The particle filter integrates both process and measurement models and allows non-linear models and non-gaussian probability distributions of the states with possibly multiple modes.

Section 2 introduces related work on the topic which has been used to define the models or has given ideas to create them. Section 3 gets into the details of the methods used for the implementation of the particle filter and the definition and training of the models. Section 4 shows and analyses the results of the index prediction with real market data. Section 5 draws some conclusions and suggest new ideas for future work.

2 Related Work

Capital Assets Pricing Model introduced by Treynor, Sharpe, Lintner and Mossin in the 1960s is a model that tries to explain how individual stock prices develop and how they are related to market development. It assumes return on a stock has two components, the systematic component associated to the market return and the residual, which equals zero in mean.

$$E(R_i) = \beta(E(R_m)) + E(\alpha_i) \tag{1}$$

CAPM theory is reviewed in [1]. Risk-free rate is assumed to be zero. Since the return of a stock is a relative measure for its price, this concept seems to be also applicable to the stock prices.

Assuming that the market is well represented by the index, the last might be correlated to individual stocks as well and might move according to some sets of coefficients, a weighted sum of the stocks. The idea of creating a prediction for an index based on a multiple linear regression appears in [4] although using earlier day's information for the one-day prediction instead of information of its constituents on the day. Via least squares multiple regression a random variable can be explained through another set of random variables, in the case here the index through the small set of constituents.

In [3], stock market prices behaviour is said to be that of a random walk. This suggest that future behaviour of stock market prices cannot be forecast based on its past. Therefore, that there exists no possible prediction that consistently gives any advantage over just choosing a random number. This idea lead to the first process model in which only noise will be added to the previous state, a random walk process.

This last hypothesis has been disputed by many, as it is a fact that investors have been profiting for years from diverse strategies based on their knowledge and the study of historical data, what is known as technical analysis. Assuming this information can be obtained, one can have a estimate of how much the market is going to move up or down, with this idea the second process model is developed.

3 Methods

In this section the particle filter is introduced along with the linear multiple regression used to train the measurement model and the process models analysed.

3.1 Particle Filter

A particle filter is a non-parametric filter in which the posterior probability distribution is represented by a finite number of samples, giving an approximation of all possible distributions. Each sample is called particle and is a hypothesis of what the true state can be at each time instant. The

likelihood of a hypothesis to be included as a particle in the particle set is proportional to the posterior, so a region of the particle space which is more densely populated denotes that the true states has higher likelihood to be there.

The particles are generated from previous particle set defining the previous belief, that in the previous time instant, which are transformed with a function that models the process and gives a hypothetical state to current time. For each particle a weight or importance factor is computed in order to incorporate the measurement, assumed to be a noisy observation of the state, in the particle set as the probability of the measurement given the particle. These weights are used in what is known as the resampling or importance resampling in which new particles are drawn from the updated particle set according to their weights, so changing the distribution of the particles so it approaches the true distribution or target distribution. Particles with low weights tend to disappear from the particle set after resampling while particles with high weights tend to be duplicated as are drawn more than once (due to drawing with replacement).[5]

Different approaches to the resampling process exists. Systematic resampling is implemented in this case as it gives three advantages: covers the sample space in a systematic way, if all samples have same importance weights no particles are lost and also its implementation complexity is smaller, so it is also faster. [5]

3.2 Linear Multiple regression for measurement model

The measurement model is a weighted average of the most significant stock prices that gives an estimate index. In order to find the weights for each of the stocks a linear multiple regression is done over the training data. This gives the weight vector $\hat{w}$ that minimizes the sum of squared error at each point between the predictions with the weighted average of the stocks price vector x_i and the real values y_i of the index for the training data.

$$\hat{w} = \arg\min_{w} \sum_{i} (y_i - w^T x_i)^2 \tag{2}$$

Measurement model noise is assumed to be Gaussian when computing the likelihood of the measurement given the particles.

3.3 Process Models

With the first hypothesis that the index behaves as a random walk so there is no new information to be subtracted, the first process model applied is just adding Gaussian noise to the previous value. In the second process model it is assumed that there is actually a prediction scheme that gives a trend as a result in the way of the market is going to go up or down by an X percent. Therefore, the process is the previous value plus the percentage that is going up each day plus Gaussian noise.

4 Experiments & Results

The market data used to run the experiments has been obtained from Yahoo Finance website. The S&P500 index and its most significant constituents in terms of high weight are included. The data used is the daily close price of each stock or index. The datasets are divided into training and validation data set. The training data set contains 1194 datapoints and the validation data set contains 14 datapoints.

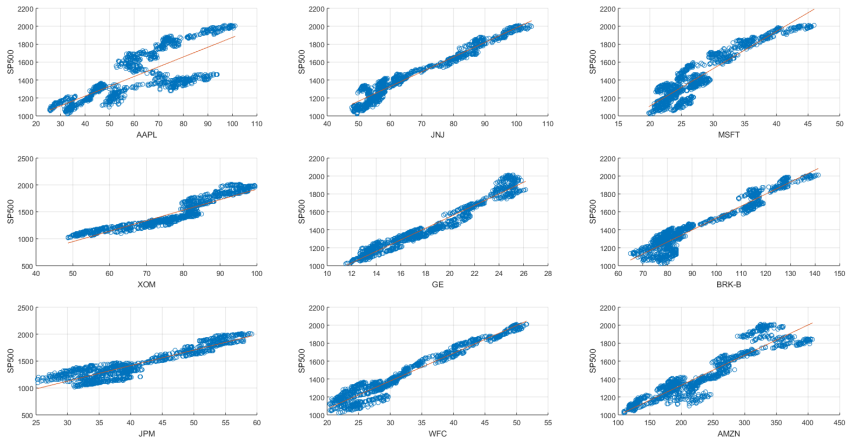

Figure 1: Scatter plot of the different stocks against the index and linear regression.

Figure 1 shows scatter plots of the different stocks against the index and Table 1 shows the correlation coefficients of each of the stock price se-

Table 1: Correlation coefficients of the stock prices against index for training and validation datasets

	AAPL	JNJ	MSFT	XOM	GE	BRK-B	JPM	WFC	AMZN
Training	0.7728	0.9751	0.9295	0.9314	0.9778	0.9572	0.9234	0.9742	0.9301
Validation	0.6452	0.9731	0.9723	0.9699	0.8916	0.8017	0.8899	0.9375	0.9661

ries against the index for the training and the validation data. Only AAPL has a significantly lower correlation coefficient, which can also be seen in the scatter plot being more spread.

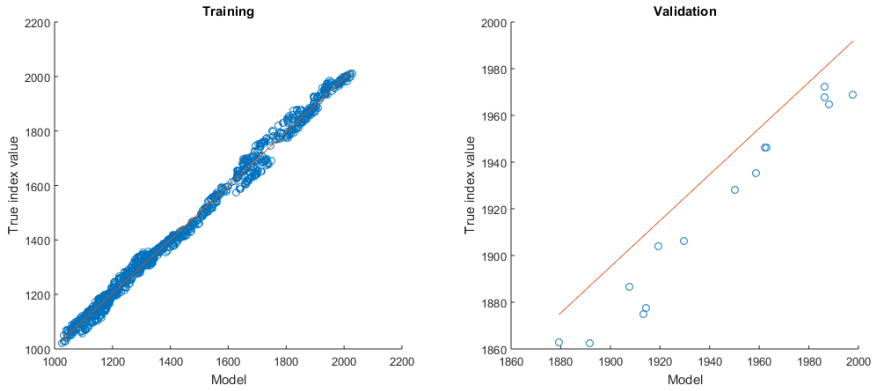

Figure 2: Scatter plot measurement model against true index value.

Figure 2 shows scatter plots of the measurement model for both training and validation data. It can be seen how for the validation data the measurement model still behaves quite good though having a small bias downwards.

The measurement model estimate gives by itself a bias of 17.5589 and a RMSE of 19.0672 on the true values of the training data, this values should be considered in order to see if there is an improvement using the particle filter and the different process models.

Figure 3 shows the resulting estimate using the first process model with only added noise in the particle filter. The number of particles is set to 2000 so it is big enough so it does not interfere with the results but it is not too computationally demanding, the same value is used in all the experiments. In some points it gets better than the measurement only but in general it cannot be said that it helps converge to the true value, clearly if

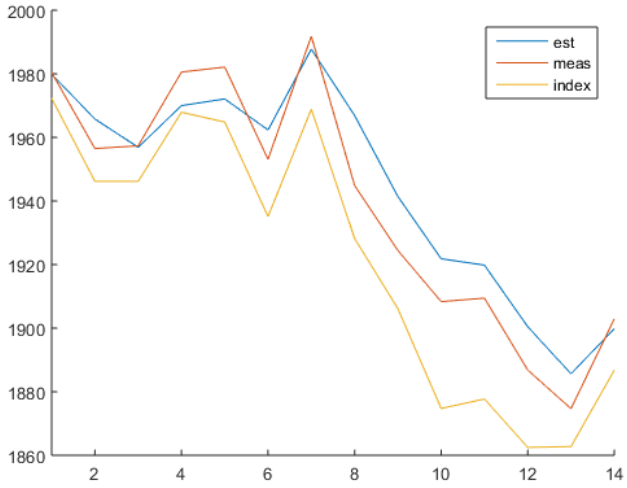

Figure 3: Measurement, estimate and true value of the index for the first process model (only noise/no prediction).

the process noise does not contain any new information the best estimate is the measurement itself. The figure is computed with the variance values for process and measurement noise that give the best results in terms of RMSE and bias for the next case in order to compare, process noise variance is 150 and measurement noise variance is 200. For this values the RMSE of the estimator is 26.5755 and the bias is 22.3099, which means it is a worse estimator than the measurement alone.

Figure 4 shows the resulting estimate using the second process model in which a trend estimate is included besides the added noise. In this case the trend estimate in the process model is somehow correct and the estimate tracks the true value of the index better than the measurement alone. Table 2 shows the RMSE and bias values of the estimate respect the true value. The RMSE and bias values for the best case drop considerably respect the previous process model.

Figure 5 shows the resulting estimate using the second process model. In this case the trend estimate used in the process model is not correct and as one can expect the estimate deviates from the measurement and has a greater bias. The RMSE and bias values increase significantly with respect to previous cases, with an RMSE value of 56.4367 clearly influenced by a

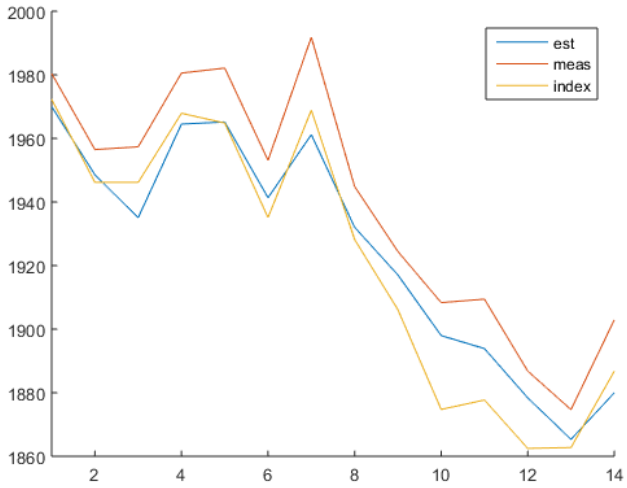

Figure 4: Measurement, estimate and true value of the index for the second process model (trend prediction plus noise).

Table 2: Table with RMSE and bias values for second process model estimate, with correct trend estimate. Measurement noise variance changes between columns and process noise variance between rows.

RMSE	100	150	200	250
100	12.9365	10.9399	13.4122	12.4415
150	13.3371	12.7350	10.1451	10.6351
200	15.9329	12.1659	12.6849	11.5774
250	15.8717	13.4899	12.7356	13.6956
Bias	100	150	200	250
100	9.9782	9.4168	10.6320	9.1596
150	9.8538	10.0985	8.0362	8.3504
200	12.9794	9.1832	9.3363	10.0723
250	13.3197	10.8767	9.9256	10.9502

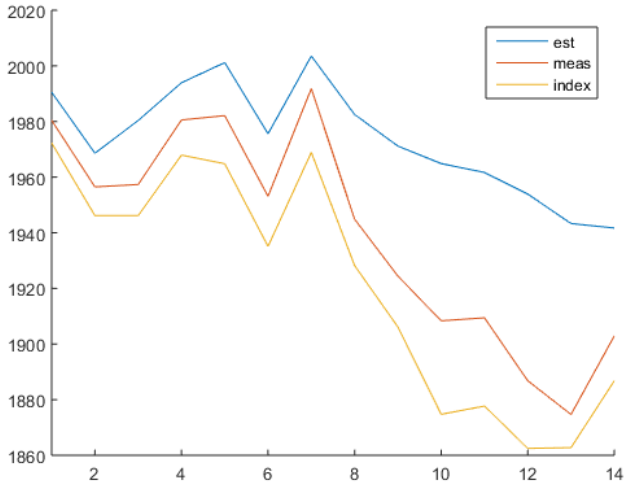

Figure 5: Measurement, estimate and true value of the index for the second process model (trend prediction plus noise), being it an incorrect prediction.

bias of 50.6296.

5 Conclusions

A particle filter approach to the prediction of stock indexes is analysed. The measurement model is a weighted average of its most significant constituents while two cases for the process model are analysed. In the first a random walk is mimicked by perturbing previous state, providing no new useful information so giving no performance gain. In the second a rough trend estimate is assumed to be available and is applied as the process model, giving considerable enhancement over only having the measurement as the estimator if the trend estimate appears to be right.

The particle filter is only useful if both the measurement model and the process model add useful and different information about how the index is going to behave. In the case of a good process model that is able to extract the real mechanics of the market the performance might not rise significantly given the fact that each of the stocks used for the measurement model already follow the market itself in some sense, so it is information

that is already in the measurement. However, given that our measurement model is composed by perfect estimates (the actual daily values which is not possible to have beforehand) results with real estimates which give a worse performing measurement model would allow more room for improvement on the process model.

References

[1] Fama, Eugene F., and Kenneth R. French. (2004) *The Capital Asset Pricing Model: Theory and Evidence.* Journal of Economic Perspectives, 18(3): 25-46.

[2] Fama, Eugene F. (1965)*Random walks in Stock Market Prices* Financial Analysts Journal (September/October):55-59.

[3] Fama, Eugene F. (1965)*The Behavior of Stock-Market Prices* The Journal of Business - Vol. 38, No. 1. University of Chicago Press.

[4] Hallgren, Fredrik (2011)*On Prediction and Filtering of Stock Index Returns* Thesis, Department of Mathematics, KTH Royal Institute of Technology.

[5] Thrun, S., Burgard, W., Fox, D. (2006) *Probabilistic Robotics*. MIT Press.