

Support Vector Regression with Lévy Distribution Kernel for Stock Forecasting

Lucas K.C. Lai, James N.K. Liu and Yanxing Hu

Department of Computer Engineering, Hong Kong Polytechnic University, Hong Kong, China
lucaskclai@yahoo.com, csnkliu@polyu.edu.hk, csyhu@comp.polyu.edu.hk

Abstract - Normal distribution has been widely applied in modern financial time series forecast. Lévy distribution is another alternative to normal distribution which this paper would like to explore. It has been demonstrated in this paper that Support Vector Regression (SVR) using Lévy distribution kernel is a robust forecasting tool and performs very well in the following experiments. Three stock Indexes are selected to test the (SVR) forecasting model. They are Hong Kong - Hang Sang Index (HSI), U.S. - Dow Jones Industrial Average Index (DJ) and China - Shanghai Composite Index (SH). It has been discovered that the SVR using Lévy kernel has given better performance in 9 out of 24 tests when compared with that of the commonly used RBF kernel. It shows promising result in general.

Index Terms - Normal distribution, Lévy distribution, Probability density function, Mercer Condition and Support Vector Regression.

I. Introduction

Normal distribution is widely used in financial model [7]. In the famous Black-Scholes [4] formula for call option price, stock price is based on Brownian motion which follows normal distribution. Robert Merton and Myron Scholes won the Nobel Prize in Economic Sciences in 1997 because of this workable option-pricing model. This famous formula also led to the collapse of Long Term Capital Management Incorporation which made a 5 Trillion US dollar hole in financial history. Subsequently, it sparked the financial crisis in 1997 which hit all Asian markets drastically. Yet, it is still widely used by options market participants even after the notorious financial tsunami in 2007. It is arguable that the financial tsunami has anything to do with the assumption that market behaves like normal distribution. But connection of the assumption to the first financial crisis has been widely recognized particularly the Black-Scholes formula stock price is simulated under normal distribution.

Estimation of parameters and testing of hypothesis are the two cornerstones of modern statistics[9]. From statisticians their point of view, all data are classified into different types of distribution using statistical analysis. Not all data are available such as the income of each individual in a society, statistical samples are taken and a parameter to represent the population such as the mean (average income of the society) is estimated from the sample using statistical inference. Here, estimation of the parameter is for present phenomena. In stock forecasting, historical samples are known and there is no need to estimate the parameter. The goal is to forecast the future movement of stock price. In this paper, it has been discovered that the historical stock movement does not follow normal distribution

rather it follows heavy tail like distribution such as Lévy distribution according to literature review [5]. The reason to employ normal distribution in stock movement is its simplicity and the characteristics of normal distribution. It is a continuous distribution with conjugate family, moment of generating function and the famous central limit theorem that a large number of independent random variables, each with a well-defined mean and well-defined variance, will be approximately normally distributed. Lévy distribution is one of the few distributions which is stable and that has probability density functions which can be expressed analytically, the others being the normal distribution and the Cauchy distribution. All three are special cases of the stable distributions that do not generally have a probability density function which can be expressed analytically.

Under Efficient market hypothesis (EMH) definition [4], US market is a strong form EMH while China market is a weak form EMH. We selected HSI because it is semi-strong form EMH. Unlike the US market which has more than 140 years of history, Hong Kong is relatively new with less than 60 years of history while being the third largest financial trading centre. Before it was a follower of the US market until recent Chinese market has significant impact on it. It is our belief that in such market, our model will perform better. Chen [3] forecasted the volatility of stock index and Olson [2] predicted the stock returns which are an indirect approach for the actual index value. The actual index value from these approaches may not be useful. In our previous paper [1], the best forecast result of the HSI 2008 first 5 days close value (based on the 15 days exponential moving average) is 0.82 mean absolute percentage error (MAPE). When it converts back to actual daily close value, the MAPE is 2.129.

The objective of this paper is to extend the work of [1] to develop a consistent approach in stock forecasting using SVR so that a platform to compare different forecasting tools can be established. The performances of SVR in predicting stock prices in the Hang Sang Index (HSI), Dow Jones Industrial Index (DJ) and Shanghai Composite Index (SH) over a 5-day and 22-day horizons respectively have been examined. The experiments are carried out in MATLAB R2011 environment with algorithm developed by the authors.

The rest of this paper is organized as follows. Section 2 describes SVR, Normal distribution and Lévy distribution. Section 3 describes our experiments and discusses the accuracy of those results. Section 4 offers our conclusion and outlines the future work.

II. Support Vector Regression

The objective of this paper is to investigate the advantage of predicting the 5-day and 22-day horizons of HSI, DJ and SH values given the historical values of the 3 indexes using SVR model and the MAPE value is below 2. In 2001[6], the winning team of power loading competition by EUNITE got their MAPE as low as 1.95. Inspired by this competition, a forecasting MAPE target result is set. The historical data of HSI, DJ and SH values between the year 2002 and 2007 were downloaded from the Yahoo financial website. They are organized in four datasets. The first two sets, corresponding to year 2006 and years 2002 to 2006 as training samples will be used to predict the values in January of year 2007. The third and fourth sets, corresponding to year 2007 and years 2003 to 2007 as testing samples to predict the values in January of year 2006.

A. Support Vector Regression

The following is a brief description on SVR [7, 8] for nonlinear function estimation such as the financial times series. In the primal weight space the model takes the form

$$f(x) = \omega^T \phi(x) + b, \quad (1)$$

with given training data $\{x_k, y_k\}_{k=1}^N$ and $\phi(\cdot): R^n \rightarrow R^{nh}$ a mapping to a high dimensional feature space which can be infinite dimensional and is only implicitly defined. Note that in this nonlinear case the vector ω can also become infinite dimensional. The optimization problem in the primal weight space becomes

$$\min_{w, b, \xi, \xi^*} J_p(\omega, \xi_i, \xi_i^*) = \frac{1}{2} \omega^T \omega + C \sum_{i=1}^l (\xi_i + \xi_i^*) \quad (2)$$

subject to:

$$y_k - \omega^T \phi(x_k) - b \leq \varepsilon + \xi_k, \quad k = 1, \dots, N$$

$$\omega^T \phi(x_k) + b - y_k \leq \varepsilon + \xi_k^*, \quad k = 1, \dots, N$$

$$\xi_i, \xi_i^* \geq 0, \quad k=1, \dots, N,$$

Applying the Lagrangian and conditions for optimality, the following is the dual problem

$$\begin{aligned} \max_{\alpha, \alpha^*} J_D(\alpha, \alpha^*) = \\ \frac{1}{2} \sum_{k,l=1}^N (\alpha_k - \alpha_k^*)(\alpha_l - \alpha_l^*) - \varepsilon \sum_{k=1}^N (\alpha_k - \alpha_k^*) + \sum_{k=1}^N y_k (\alpha_k - \alpha_k^*) \end{aligned} \quad (3)$$

subject to :

$$\sum_{k=1}^N (\alpha_k - \alpha_k^*) = 0$$

$$\alpha_k, \alpha_k^* \in [0, c]$$

Here the kernel trick has been applied with $K(x_k, x_l) = \phi(x_k)^T \phi(x_l)$ for $k, l = 1, \dots, N$. The dual representation of the model becomes

$$f(x) = \sum_{k=1}^N (\alpha_k - \alpha_k^*) K(x, x_k) + b \quad (4)$$

Consider the following Vapnik's ε -insensitive loss function

$$L_\varepsilon(y - f(x)) = \begin{cases} 0, & \text{if } |y - f(x)| \leq \varepsilon \\ L(y - f(x)) - \varepsilon & \text{otherwise} \end{cases} \quad (5)$$

Eq.5 is a convex cost function where $L(\cdot)$ is convex.

Primal problem

$$\min_{w, b, \varepsilon, \varepsilon^*} \frac{1}{2} w^T w + C \sum_{k=1}^N (L(\varepsilon_k) + L(\varepsilon_k^*)) \quad (6)$$

subject to

$$y_k - w^T \phi(x_k) - b \leq \varepsilon + \varepsilon_k$$

$$w^T \phi(x_k) + b - y_k \leq \varepsilon + \varepsilon_k^*$$

$$\varepsilon_k, \varepsilon_k^* \geq 0$$

where $\varepsilon_k, \varepsilon_k^*$ are slack variables. Here, x_k is mapped to a higher dimensional space by the function ϕ and ξ_k is the upper training error (ξ_k^* is the lower) subject to the ε -insensitive tube $|y_k - w^T \phi(x_k) - b| \leq \varepsilon$. The parameters which control the regression quality are the cost of error C , the width of the tube ε , and the mapping function ϕ .

The constraints imply that we should put most data x_k in the tube $|y_k - w^T \phi(x_k) - b| \leq \varepsilon$. If x_k is not in the tube, there is an error ξ_k or ξ_k^* which we must minimize the objective function SVR to avoid under-fitting or over-fitting the training data by minimizing the training error $C \sum_{k=1}^N (L(\varepsilon_k) + L(\varepsilon_k^*))$ as well

as the regularization term $\frac{1}{2} w^T w$.

The Lagrangian for this problem is

$$\begin{aligned} L(\omega, b, \varepsilon, \varepsilon^*; \alpha, \alpha^*, \eta, \eta^*) = \frac{1}{2} \omega^T \omega + \\ c \sum_{k=1}^N (L(\xi_k) + L(\xi_k^*)) - \sum_{k=1}^N \alpha_k (\varepsilon + \varepsilon_k - y_k + \omega^T \phi(x_k) + b) - \\ \sum_{k=1}^N \alpha_k^* (\varepsilon + \varepsilon_k^* + y_k - \omega^T \phi(x_k) - b) - \sum_{k=1}^N (\eta_k \varepsilon_k + \eta_k^* \varepsilon_k^*) \end{aligned} \quad (7)$$

With Lagrange multipliers $\alpha_k, \alpha_k^*, \eta_k, \eta_k^* \geq 0$ for $k=1, \dots, N$.

Dual problem

$$\max_{\alpha, \alpha^*, \eta, \eta^*} J_D(\alpha, \alpha^*, \eta, \eta^*) \quad (8)$$

subject to

$$\sum_{k=1}^N (\alpha_k - \alpha_k^*) = 0$$

$$cL'(\varepsilon_k) - \alpha_k - \eta_k = 0, k=1, \dots, N$$

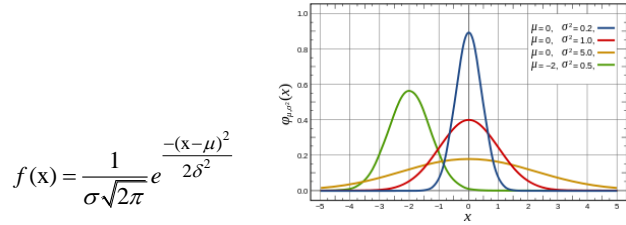
$$cL'(\varepsilon_k^*) - \alpha_k^* - \eta_k^* = 0, k=1, \dots, N$$

$$\alpha_k, \alpha_k^*, \eta_k, \eta_k^* \geq 0, k=1, \dots, N.$$

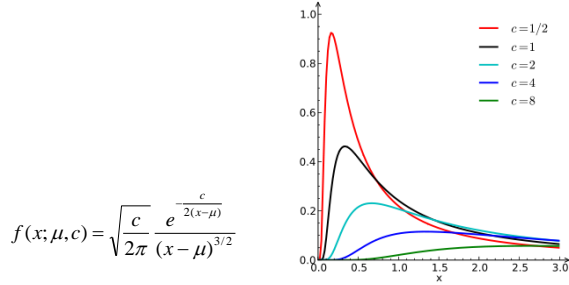
So far from (1) to (8), SVR estimation function combined with the loss function is the foundation of the SVR.

B. Normal and Lévy Distribution

The probability density function (pdf) of the distribution is used to simulate the stock movement in financial forecasting. To use pdf in (4) of the SVR, it must satisfy the Mercer condition and it should be positive definite. Normal distribution is the good candidate for reasons mentioned above with the following pdf equation and figure.



As one of the few stable distributions, Lévy distribution also satisfies Mercer Condition with the following pdf equation and figure.



III. Experimental and results

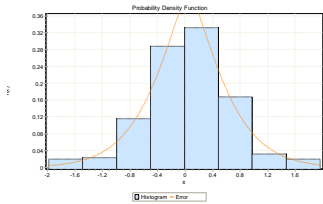


Fig. 1 DJ 2006 data set

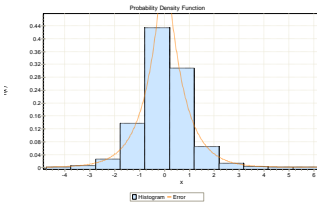


Fig.2 DJ 2002-6 data set

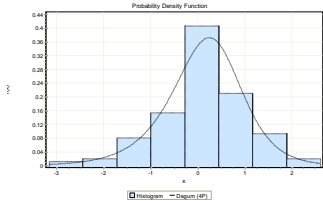


Fig.3. HSI 2006 data set

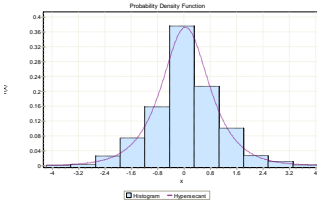


Fig.4. HSI 2002-6 data set

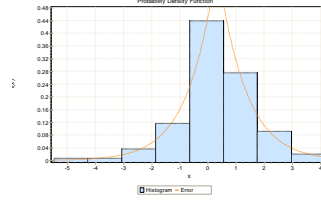


Fig.5 SH 2006 data set

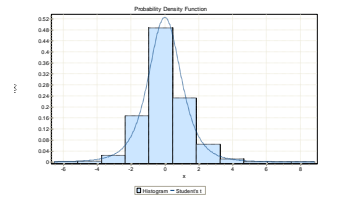


Fig. 6 SH 2002-6 data set

Fig.1 to Fig.6 are the histograms of the log return for the dataset 2006 and 2002-6 of the three different markets. All of the above histograms have kurtosis more than 3 which means they are not normally distributed. The difference between Fig.1 to Fig.6 compared with the normal distribution curve is the heavy tail behavior.

As explained in the introduction, SVR maps data into higher dimensional space using kernel function. In (4) RBF kernel function is $K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$. The formula in

that kernel function is $e^{(-\text{gamma}^*(x_k - x_l)^2)}$ which is a normal pdf. Another kernel using Lévy pdf with the formula $\frac{e^{(-\text{gamma}^*(x_k - x_l))}}{(x_k - x_l)^{3/2}}$ has been included in SVR model. Table I is

the result from SVR using normal distribution kernel. The cost constant C as in (2) has different settings. The higher the C, the less the percentage of error. The g which stands for the gamma in the kernel function also has different settings. The value of g is to adjust the kernel function and the higher the value the greater the mapping dimension is. The parameters C and g in this paper were based on the previous work by [1].

TABLE I Prediction Result of SVR using Normal Distribution

Dataset range	Forecast Horizon	SVR DOW	SVR Hang Sang	SVR SH
2006	2007 5 days	0.38	2.71	9.74
2002 to 2006	2007 5 days	0.70	2.71	7.32
2007	2008 5 days	1.94	1.75	1.85
2003 to 2007	2008 5 days	1.14	0.94	1.85
2006	2007 22 days	1.22	5.58	19.24
2002 to 2006	2007 22 days	1.22	5.58	18.23
2007	2008 22 days	4.48	5.07	8.53
2003 to 2007	2008 22 days	2.86	5.52	8.60

Short term forecast for 2007 with 5-days horizon considers the 2006 dataset only as input. The lowest MAPE result is 0.38 in DOW Index. For 2007 with 5-day forecast horizon using a dataset from 2002 to 2006, the lowest result is 0.70 also in DOW Index. Long term forecast for 2007 with 22-day horizon considers the 2006 dataset only as input, the lowest result is 1.22 in Dow Index. For 2007 with 22-day

forecast horizon using a dataset from 2002 to 2006, the lowest result is 1.22 in DOW Index.

Short term forecast for 2008 with 5-days horizon considers the 2007 dataset only as input. The lowest MAPE result is 1.75 in Hang Sang Index. For 2008 with 5-day forecast horizon using a dataset from 2003 to 2007, the lowest result is 0.94 also in Hang Sang Index. Long term forecast for 2008 with 22-day horizon considers the 2007 dataset only as input, the lowest result is 4.48 in Dow Jones Index. For 2008 with 22-day forecast horizon using a dataset from 2003 to 2007, the lowest result is 2.86 also in Dow Jones Index. In Table I, there are 10 MAPE values out of 24 lower than 2.

SVR requires much effort to tune the parameters c and g in order to get a better result. It is rather difficult to determine which setting is correct. We discovered from the experiments that the parameter c has to be set with the range between 1,000 and 8000 in order to produce meaningful results and the c value must be more than 500.

TABLE II Prediction Result of SVR using Lévy Distribution

Dataset range	Forecast Horizon	SVR DOW	SVR Hang Sang	SVR SH
2006	2007 5 days	0.36	7.13	4.14
2002 to 2006	2007 5 days	2.69	3.22	9.88
2007	2008 5 days	1.94	1.60	1.98
2003 to 2007	2008 5 days	1.24	1.15	21.70
2006	2007 22 days	0.98	7.15	6.49
2002 to 2006	2007 22 days	3.29	3.30	12.07
2007	2008 22 days	4.48	5.45	6.71
2003 to 2007	2008 22 days	3.21	5.36	28.76

Tables II gives improvement using Lévy distribution in the 3 markets. There are 2 in DOW market, 3 in Hang Sang and 4 in Shanghai. It seems the more efficient the market, the less Lévy distribution kernel can improve the forecasting result. The significant improvement using Lévy distribution in Shanghai market is remarkable. It is a weak-form EMH market and it is not likely to follow normal distribution movement in its price trend. This discovery is important as most of the literature review did not find any improvement using different kernels in SVR. One of the advantages of SVR is the mapping of data into higher dimension using kernel and avoid finding the distribution parameters of the original data. The kurtosis of the log return dataset DJ 2006, DJ2007, SH2006, SH2002-6 and SH2007 are 4.24, 4.56, 5.6, 7.03 and 4.95 respectively. After the dataset are mapped into higher dimension using normal distribution, the kurtosis of the log return dataset DJ 2006, DJ2007, SH2006, SH2002-6 and SH2007 are 6.25, 10.63, 21.92, 117.23 and 6.36 respectively. After the dataset are mapped into higher dimension using Lévy distribution kernel, the log return kurtosis reduces to 4.09, 4.54, 6.17, 7.94 and 4.35 respectively. It seems the closer the kurtosis value of

the higher dimension data to the original dataset, the better is the forecasting result.

IV. Future Work and Conclusion

This paper has demonstrated the potential to use Levy distribution in SVR model. The improvement in forecasting accuracy is significant in the 3 markets especially in Shanghai market. We believe this is a significant contribution in forecasting methodology. Our finding shows that the log return kurtosis value in higher dimension is key to compare the result.

It is important to point out that the experiments have demonstrated promising forecasting result in strong form EMH market such as DOW in USA especially using SVR model. 6 out of 8 forecasting values have MAPE under 2 which is the objective of this paper. For HSI Hong Kong market, the performance is similar to DOW despite it is a semi-strong EMH market but it can still provide a good decision making platform for investment. For weak form EMH SH in China, the forecasting ability is obviously poor. This is because the volatility of the market strongly affects the forecasting power. In future, it is necessary first to forecast the volatility of the market and then develop a hybrid SVR model to input the volatility attribute into it in order to improve the performance.

There are a few more distribution functions that can be employed as SVR kernel such as Hyperbolic Secant, Student's t and Laplace distribution which the authors would like to explore their capabilities with more testing in future work.

Acknowledgement

The authors would like to acknowledge the partial support of the CRG grant G-YM07 of The Hong Kong Polytechnic University.

References

- [1] L.K.C. Lai and J.N.K. Liu, "Stock forecasting using support vector machine", *Journal of Machine Learning and Cybernetics (ICMLC)*, pp1607-1614, 2010.
- [2] D. Olson and C. Mossman C, "Cross-correlations and Predictability of Stock Returns," *Journal of Forecasting J. Forecast.* 20, 145-160 (2001).
- [3] C Chen, W.K. Hardle and K. Jeong, "Forecasting Volatility with SVM-Based GARCH Model," *Journal of Forecasting, J. Forecast.* 29, 406-433(2010), DOI:10.1002/for.1134.
- [4] Z. Bodie, A. Kane and A. J. Marcus, *Investments Sixth Edition*, Mcgraw-Hill International Edition 2005 ISBN 0-07-286178-9.
- [5] C. Tu, M. Clyde and R.L.Wolpert, "Lévy Adaptive Regression Kernels," *Duke University Department of Statistical Science*, 2007.
- [6] C.J. Lin, B.J. Chen and M.W. Chang, "Load Forecasting Using Support Vector Machines: A Study on ENUNITE Competition 2001," *Department of Computer Science and Information Engineering*, National Taiwan University, 2001.
- [7] V Vapnik, SE Golowich, A Smola, *Support vector method for function approximation, regression estimation, and signal processing* *Advances in neural information processing systems*, pp.281-287,1997
- [8] Y.X. Hu and J.N.K. Liu, "Application of Feature Weighted Support Vector Regression Using Grey Correlation Degree to Stock Price Forecasting", *Neural Computing and Applications, Neural Computing and Applications*, Volume 22, Issue 1, 143-152 (2013).
- [9] Y.L. He, J.N.K. Liu, X.Z. Wang and Y.X. Hu Optimal bandwidth selection for re-substitution entropy estimation, *Applied Mathematics and Computation*, Volume 219, Issue 8, 3425-3460 (2012).