Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 14 (16): 1883-1888, 2014
ISSN 1812-5654 / DOI: 10.3923/jas.2014.1883.1888
© 2014 Asian Network for Scientific Information

Comparison of Classification Methods for Predicting the Movement Direction of Saudi Stock Exchange Index

[1]M. Alrasheedi and [2]A. Alghamdi
[1]Department of Quantitative Methods, School of Business, King Faisal University,
Hofuf Al-Ahsa, Saudi Arabia
[2]King Abdulaziz University, Jeddah, Saudi Arabia

Abstract: Stock exchange index plays an important role in the financial structure of a country and control the up and own movement of price structures of various indexes for business decisions. It is important for business in the stock markets to have an idea about future stock price development. The main objective of this study was to compare the different classification methods for predicting the movement direction of Saudi stock exchange index-TASI (Tadawul All Share Index]. The study applied four classification methods namely Linear Discriminant Analysis (LDA), Logistic Regression (LR), Support Vector Machines (SVM) and Least Squares Support Vector Machines (LS-SVM) as classification methods to predict the movement direction (up or down) of Saudi Stock Exchange Index (TASI) called as Tadawul All Share Index from Aug 2006 to April 2013. The study included several publicly available financial data such as Dow Jones Index, closing price, opening price, SABIC trading volume, crude oil price and TASI as the independent variables. The study chose 5-fold cross validation out-of-sample hit rates to compare the performance. For all the methods, except LS-SVM, hit rates ranged between 61 and 63 and only slightly lower for a reduced set of independent variables . With a data set from April 2009 to April 2013, the hit rates were about 2% higher than the normal rates. Overall, the LR, LDA and SVMs performed well for the price index data used in the study.

Key words: TASI, linear discriminant analysis, logistic regression, support vector machines, k-fold cross validation

INTRODUCTION

Prediction of future stock prices is very difficult as these are influenced by many factors and do not have a simple structure. Nevertheless, the predictions are very important for business decisions. Oftenly, it is helpful to know the direction of the movement (up or down) which is an easier task to predict, but still difficult for financial data. Also, information about future up or down movement can be used for hedging strategies. For example, the researchers decided to use the classical methods such as Linear Discriminant Analysis (LDA) (Venables and Ripley, 2002) and Logistic Regression (LR) (Hosmer and Lemeshow, 2000) to predict the future movement direction and compare them with the latest methods such as support vector machines (Cortes and Vapnik, 1995) and least squares support vector machines (Suykens and Vandewalle, 1999). All these classification methods were used in the president study to distinguish between the two stock price movement directions. The first two methods are linear and the results are easy to interpret, whereas the latter two are non-linear (linear in a high-dimensional space) and more flexible for the classification task. The investigators used the out-of-sample hit rate for comparison between the two methods. The results were verified with 5-fold cross validation to minimize the effect of the division into training and testing set. There are more sensitive measures for the efficiency of the methods (Pohar *et al*., 2004), but with respect to its practical application, the number of correct predictions is most important. This study chose to apply these methods to available daily data from Saudi Stock Exchange (SSE). Recently, Alrasheedi (2012) used the first two methods for predicting the SABIC movement direction. The review of literature indicated that not much research has been conducted on this topic in Saudi Arabia. Therefore, the main objective of was to expand this approach with more and refined methods for predicting the direction movement of Saudi stock exchange.

RESEARCH METHODOLOGY

Data selection: The study included several publicly available financial data to predict the up or down movement of the Saudi Stock Exchange (SSE) also known as Tadawul All Share Index (TASI) obtained from

Corresponding Author: M. Alrasheedi, King Faisal Universsity, Hofuf Al-Ahsa, Saudi Arabia

http://www.investing.com/indices/tasi-historical-data. Since, it is the biggest company in the Saudi Stock Exchange market, the researchers included Saudi Basic Industries Corporation (SABIC) values for opening, low, high, volume, turnover and number of trades obtained from (http://www.tadawul.com.sa). The study also included the Dow Jones Index (DJI) as an indicator of influences from the global market by the source http://uk.finance.yahoo.com (3). As the Saudi economy is mainly oil based, therefore 11 versions of oil prices (crude oil, gasoline, diesel, propan, etc.) were downloaded from the US Energy Information Administration (http://www.eia.gov/dnav/pet/pet_pri_spt_s1_d.htm). The time span chosen was from 1st Aug 2006 to 1st April 2013. All the data were transformed into log returns, i.e., if X_t is the time series, then the log returns are log (X_t/X_{t-1}). This transformation removed the trend patterns and has other desirable qualities (Fama, 1965). The direction of TASI movement was then indicated by the sign of the log return i.e., positive means up and negative means down. The study chose to include 0 into the down movement. Different Stock Exchanges have different trading days. For example DJI usually works from Monday to Friday and the SSE from Saturday to Wednesday. The study tried two approaches to synchronize the data: First to omit all the days without full data set thus leaving only 622 days. Secondly, to fill in the last value of the time series each time there is a gap (last value carried forward) which gives zeros in the log returns resulting in 1669 days. With such a big number of independent variables which might be quite similar, there is the risk of multicollinearity. In the worst case (strict) multicollinearity means that the applied statistical methods have numerical problems and will not be able to produce a result. But nonetheless, so many variables make the model more complicated and might not add more information. However, effort was made to check for correlations between the log returns of all independent variables for both approaches to synchronize the different trading days. It was decided to keep only a set of independent variables, where none of the mutual correlations is bigger than 0.4 (Yuan, 2011) which is, when we want to predict tomorrow's TASI movement i.e., the closing price for TASI, SABIC volume, opening price for DJI, closing price for DJI, DJI volume, price for crude oil (RWTC) all from today and additionally yesterday's closing price for TASI. It was felt that any value from tomorrow, like the TASI high or low should not be included, because there is no clear cut off point in time until which we gather information. For a brief check we included it at the end to compare with other results (Ou and Wang, 2009).

The study applied all the methods to a reduced set of independent variables, where the SABIC volume and DJI volume were omitted. The chosen time span covers the economical crisis which started in December 2007. The graphs of the time series show different behaviour before and after the crisis: For example TASI seems to go into opposite directions in comparison to DJI before December 2007 and later they seem to be more parallel (Fig. 1). Therefore, the study applied all the methods for the time period from 1st April 2009-2013 only (375 trading days with complete data).

The study applied a number of different methods to forecast the movement direction of TASI. First, the classical methods linear discriminant analysis and logistic regression were chosen and then compared with more modern approaches such as support vector machines and least squares support vector machines. All the analyses were done with the software R (regression). For the evaluation of all methods, the so-called k-fold cross validation was used, where the full data set was divided randomly into k (in this case 5) subsets. Out of these, four subsets were used together to train the model and the fifth was used for validation. The results for all the five validation subsets were compiled together to give the final out-of-sample hit rate. Since the division into subsets was random, the final hit rate might differ if the *k*-fold cross validation is run again.

Linear discriminant analysis (LDA): The Linear Discriminant Analysis (LDA) was applied to divide the feature space (spanned by the independent variables) into two parts, in this study for the movement directions up or down. The division was done in a linear way by a hyperplane which means estimating parameters for each independent variable so that the linear combination of the

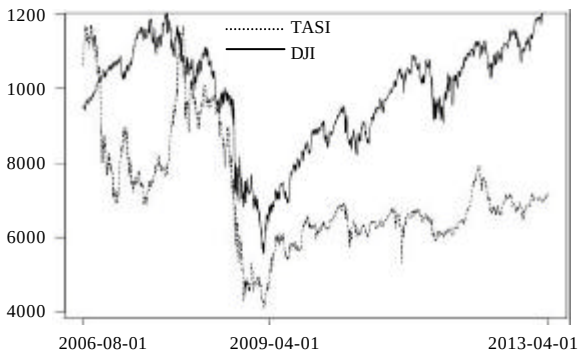

Fig. 1: Prices of TASI and DJI (rescaled by 85%, from 2006-2013

parameters and the variables gives an equation which can be used to predict in which class (0 or 1) a new observation is likely to be found. It was assumed that the independent variables X_i are normally distributed with means μ_0 and μ_1 for class 0 and 1, respectively. Under the additional assumption that the covariances (σ) of the two classes are identical and have full rank, the optimal Bayes solution for predicting that point is from the second class is:

$$w \bullet x < c$$

for some threshold constant c, where:

$$w = \sigma^{-1}(\mu_1 - \mu_0)$$

This analysis was done in R with the function LDA from the package MASS given by Venables and Ripley (2002) and Ripley (1996).

Logistic Regression (LR): The Logistic Regression (LR) gives the odds for a certain outcome, i.e., up or down movement in our case. The log odds of the outcome is modelled as a linear combination of the independent variables X_i. The independent variables do not have to follow a certain distribution. The probability p_1 belonging to class 1 is modelled as:

$$p_1 = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

where, α and β are the regression parameters. In R the logistic regression can be done via the function glm with parameter 'family="logit"' (Hosmer and Lemeshow, 2000; Long, 1997; Venables and Ripley, 2002).

Support Vector Machines (SVM): Previously, the Support Vector Machines (SVM) were proposed for classification of movement direction of stock exchange by Boser *et al.* (1992) and Cortes and Vapnik (1995). Given a training set of instance-label pairs $(x_i;y_i)$, i = 1, ..., l, where x is n-dimensional and y can be 1 or -1, the Support Vector Machines (SVM) require the solution of the following optimization problem:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + C\sum_{i=1}^{l} \zeta_i$$

subject to $Y_i(w^T\varphi(x_i)+b) \geq 1-\zeta_i, \zeta_i \geq 0$.

C >0 is the penalty parameter of the error term. Here, the classification problem is mapped into a high-dimensional space with the help of a kernel:

$$\varphi(x_i)^T \varphi(x_j)$$

solved there with a linear separating hyperplane and mapped back again which results in a non-linear classification in the original space. The method is broadly applicable, but careful tuning of the necessary parameters is essential to obtain good results. First the kernel type has to be chosen and then the cost parameter C and possible kernel dependent parameters have to be optimised. The R package e1071 offers the function SVM for actual classification and the function tune.SVM is to perform a grid search over the parameter space. This study followed the guidance of Hsu *et al.* (2003).

Least Squares Support Vector Machines (LS-SVM): A new version of support vector machines (LS-SVM) with a least squares loss function was proposed by Suykens and Vandewalle (1999). Here, the function was minimised and takes the form:

$$\min_{w,b,\zeta} \frac{1}{2} w^T w + \frac{1}{2}\gamma\sum_{i=1}^{l} \zeta_i^2$$

Also here a careful tuning of the parameters is necessary.
A function LS-SVM can be found in the R package kernlab.

RESULTS AND DISCUSSION

By applying LDA to trading days with complete data (622 days), the following parameter estimates were obtained to predict the TASI movement at time t+1: the closing price of TASI was 20.4, SABIC volume was 0.026, opening price of DJI was 2.4, closing price of DJI was 51.8, DJI volume of 0.9, price for crude oil (RWTC) as 16.2 all at time t and closing price of TASI at time t-1 as 8.8 (all the variables were log returns). Several runs of the 5-folds cross validation gave out-of-sample hit rates between 61.4 and 63.2%. The most influential independent variables seem to be the closing price of DJI and TASI along with the crude oil price.

For LDA with the imputed data set (1669 days), a hit rate of 55.4% was obtained and the parameter estimates were difficult to interpret. An essential assumption for LDA is the normality of all the independent variables. The analysis of Q-Q-Plots and the tails of all variables deviated strongly from the middle line, indicating a

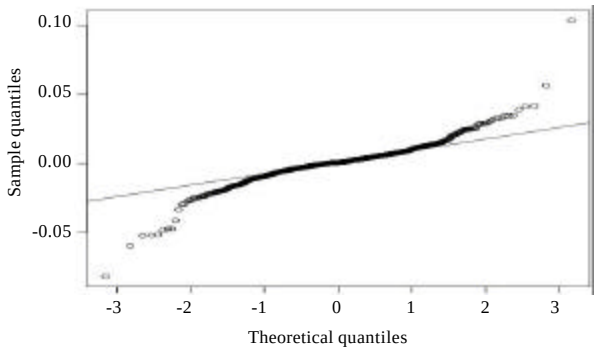

Fig. 2: Q-Q-Plot vs. normal distribution of DJI

heavy-tailed distribution of the variables (Fig. 2). The fact, that log returns of financial data are often heavy tailed and already known (Fama, 1965) and Alkhathlan and Prabakaran (2009) for TASI).

As the volumes of SABIC and DJI variables got very low estimates for the loadings thus indicating that they might not contain much information for the prediction of TASI movement. Therefore, the authors tried a reduced model without these two variables and got loadings for the closing price of TASI as 20.6, opening price of DJI as 1.55, closing price of DJI as 52.3, price for crude oil (RWTC) as 16.6 at time t and closing price of TASI at time t-1 as 8.9. The out-of-sample hit rate ranged between 61.1 and 62.7% which means that there was not much loss in performance in comparison to the full model and the advantage of a simplified model. A logistic regression was performed with the same data sets used for LDA. For the data set without missing values, the parameter estimates were: The closing price of TASI as 12, SABIC volume of 0.03, opening price of DJI as 1.7, closing price of DJI as 32.1, DJI volume of 0.57, price of crude oil (RWTC) was 9.1 at time t, closing price of TASI at time t-1 was 4.6 and the intercept as 0.32. However, at 5% level of confidence, the hypothesis cannot be rejected as the parameter could be 0 for most variables, only for the intercept, closing price of DJI and the price of crude oil. With 5-fold cross validation, the out-of-sample hit rates ranged between 61.3 and 62.4%. Also, without the volume variables, the study got estimates of closing price of TASI as 11.5, opening price of DJI as 1.2, closing price of DJI as 29.7, price of crude oil (RWTC) was 8.7 at time *t,* closing price for TASI at time t-1 was 4.5 and intercept was equal to 0.33. Again at the 5% level we can only reject the hypothesis, that the parameter could be 0, for the intercept, closing price for DJI and the price for crude oil. The out-of-sample hit rates were between 61.8 and 62.4%.

Like LDA, the most influential independent variables seem to be the closing price of DJI and TASI and the crude oil price. Also all out-of-sample hit rates were similar for LR and LDA and full or reduced set of variables. For the data set with replaced missing values again, the parameter estimates were difficult to interpret and the out-of-sample hit rates were 54.8% or worse.

However, if you apply support vector machines in a naïve way without careful parameter selection, the results are visibly suboptimal. With preselected parameters, the 5-fold cross validation out-of-sample hit rate was only 58.8%. For the tuning (selection of good parameters), the study followed the guidance of Hsu *et al.* (2003). They suggested to perform a grid search for all the parameters after choosing a kernel. The authors tried several kernels and got the best results for the radial basis function. With this choice of a kernel, the parameter γ and the cost parameter C have to be optimised. We ran models for C in $2^{-5}, 2^{-3}, \ldots, 2^{15}$ and γ in $2^{-15}, 2^{-13}, \ldots, 2^{3}$ which gave C = 2 and $\gamma = 2^{-5}$ as the best parameters. The out-of-sample hit rate for these parameters was 61.6%. A refined parameter search was made around these values for C in $2^{-1}, 2^{-0.75}, \ldots, 2^{3}$ and γ in $2^{-7}, 2^{-6.75}, \ldots, 2^{-3}$. Hsu *et al.* (2003) stated that this procedure usually gives reasonable results, but does not guarantee to find the optimum. Now the best parameters were C = 8 and $\gamma = 2^{-5.25}$. The out-of-sample hit rate for these parameters ranged between 61.6 and 63.7%. In addition to that, the trials with other kernels or other areas for the parameters or v-classification did not improve the result.

Also, the study did a similar grid search for the reduced set of variables and got $C = 2^{2.5}$ and $\gamma = 2^{-5.75}$. The out-of-sample hit rate ranged between 60.6 and 61.9%. With both sets of variables, the number of chosen support vectors was usually close to the maximum. For the data set with replacement for missing values, an out-of-sample hit rate was found to be 54.6%.

Least Squares Support Vector Machines (LS-SVM) are a variation of the classical support vector machines, but with a different loss function. It was found that the 5-fold cross validation out-of-sample hit rate varied extremely. For example, for the same parameters with four runs, the hit rates ranged between 41.8 and 52.7%. Therefore, it was not possible to select the parameters in a good way. The best hit rate obtained during the experiments was 58.8%.

It was also noticed that the behaviour of the variables in relation to each other changed over time. For example: Before the economical crisis, TASI and DJI seem to always do the opposite, but after the crash they seem to be more parallel. So, for the comparison of the models, it was decided to run all analyses also for the time span 1st April 2009-2013. The loadings for LDA were: The closing price of TASI as 0.11, SABIC volume of 0.11,

opening price of DJI as 8.9, closing price of DJI as 81.5, DJI volume of 0.74, price of crude oil (RWTC) as 21.9 at time t and closing price of TASI at time t-1as 0.39. The out-of-sample hit rates ranged between 64.3 and 67.2%.

For the logistic regression, the parameter estimates became: The closing price of TASI as -0.3, SABIC volume of 0.083, opening price of DJI as 5.97, closing price of DJI as 57.6, DJI volume of 0.57, price of crude oil (RWTC) as 14.2 at time t, closing price of TASI at time t-1 as 0.48 and the intercept was 0.46. On the 5% level, the hypothesis can not be rejected, that the parameter could be 0 for most variables, only for the intercept, closing price of DJI and the price of crude oil. With 5-fold cross validation, the out-of-sample hit rates ranged between 64.5 and 66.7%.

This study did the same two-step parameter grid search for SVM. The best parameters were C = 512 and $\gamma = 2^{-9}$. The out-of-sample hit rate for these parameters was between 65.3 and 67.2%. For a small data set starting from 27th August 2011 until 1st April 2013 (142 complete trading days), the study included the TASI high, low and opening of the day to be predicted. It was noticed that, while the hit rates were obtained around 61 and 67%, while in other studies (Ou and Wang, 2009), the hit rates for similar data were around 80-86%. With these new variables, the out-of-sample hit rates ranged between 78.2 and 80.3% for LDA with the new variables being the most influential ones. Logistic regression gave a similar picture with hit rates ranging between 75.4 and 78.9 and the SVM between 76.1 and 82.4%. A summary of all the hit rates is presented in Table 1-4.

The LDA, LR and SVM gave similar results in terms of out-of-sample hit rates. Theoretically, LDA needs all the independent variables to be normal, a debatable assumptions for our data with the heavy tails, so the application of LDA needs to be considered with care as there is no theoretical justification to guarantee good results.

However, out of these three parameters, the SVMs are the most flexible. Because, it is a non-linear method which enables it probably to give better results than the other methods. But with the data available for the present study, this is not the case. Also, It requires more experience to obtain good results, because without careful selection of parameters, the results can easily be worse than the other methods. Also the result (a set of support vectors) is more difficult to interpret than the loadings or estimated parameters of LDA and LR.

The other limitation was that LS-SVMs could not be used properly. Because the out-of-sample hit rates were highly dependent on the division of the data set. This seems to indicate over-fitting of the training set, but we could not find parameters, where it did not happen. May be this data set was more difficult for SVM and LS-SVM as the SVM was always choosing a big number of support vectors which means it has difficulties with the classification task.

Table 1: Five fold cross validation out-of-sample hit rates for Linear Discriminant Analysis (LDA), Logistic Regression (LR) and Support Vector Machines (SVM) (for least Squares Support Vector Machines (LS-SVM)

Parameters	Percentage (%)
LDA	61.4-63.2
LR	61.3-62.4
SVM	61.6-63.7
(LS-SVM)	(58.8)

We could not find good parameters) only for trading days without any missing values (622 days)

Table 2: Five fold cross validation out-of-sample hit rates for Linear Discriminant Analysis (LDA), Logistic Regression (LR) and Support Vector Machines (SVM) only for trading days without any missing values (622 days), but without the variables SABIC and DJI volume

Parameters	Percentage (%)
LDA	61.1-62.7
LR	61.8-62.4
SVM	60.6-61.9

Table 3: Five fold cross validation out-of-sample hit rates for Linear Discriminant Analysis (LDA), Logistic Regression (LR) and Support Vector Machines (SVM) where any missing values in the prices have been replaced by the last known value

Parameters	Percentage (%)
LDA	55.4
LR	54.8
SVM	54.6

Table 4: Five fold cross validation out-of-sample hit rates for Linear Discriminant Analysis (LDA), Logistic Regression (LR) and Support Vector Machines (SVM) for data from 1st April 2009-2013

Parameters	Percentage (%)
LDA	64.3-67.2
LR	64.5-66.7
SVM	65.3-67.2

In order to get better hit rates, more informative independent variables are needed. The amount of information contained in the independent variables is not sufficient for a better classification. This can also be seen from the fact, that the hit rates in the classification without the volume variables are almost identical as with all variables. Since more variables have small loadings or estimated parameters, the set of variables could probably even be reduced more without a big loss.

More information about the dependent variable is contained in the variables such as TASI high, low and open from the day for which the prediction is to be made. The brief test showed hit rates close to 80% or more. But it was observed that the high or low hit rates could be reached only with the closing price (the one to be predicted) and should not be included.

The results for the data set, where the last known value was used as replacement for the missing values, were nor satisfying. The replacement values were mostly zeros in the log returns and that might have disturbed the classification process more than helping to increase the number of trading days. When the data was used only from 1st April 2009 onwards (after the crisis), the hit rates were better. The economical situation did not change so much during that time in comparison to the earlier period. As long as the same economical situation was encountered, the model based on the later data set performed better, but it is less robust in case of another crisis.

Overall, the LR, LDA and SVMs performed well for this type of data. But, if the performance need to be improved then more informative data is needed.

CONCLUSION

The LDA, LR and SVM gave similar results in terms of out-of-sample hit rates. The out-of-sample hit rate for these parameters was between 65.3 and 67.2%. With 5-fold cross validation, the out-of-sample hit rates ranged between 64.5 and 66.7% as well as between 61.3 and 62.4% with different variables. Theoretically, LDA needs all the independent variables to be normal, a debatable assumptions for this study data with the heavy tails, so the application of LDA needs to be considered with care without theoretical justification for good results. Out of the three parameters (LDA, LR and SNM), the SVMs are the most flexible as it is a non-linear method which enables to give better results than other methods. As long as the same economical situation is encountered, the model based on the later data set performed better. Overall, the LR, LDA and SVMs performed well for this type of data. But, if the performance need to be improved then more informative and reliable data is needed.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the financial support for this project by Deanship of Scientific Research (DSR) Grant No. 140191, King Faisal University, Hofuf Al-Ahsa, Saudi Arabia.

REFERENCES

Alkhathlan, K. and S. Prabakaran, 2009. Memory effects on Saudi Arabian stock market-empirical evidence. Enterprise Risk Manage., 1: 13-34.

Alrasheedi, M., 2012. Predicting up/down direction using linear discriminant analysis and logit model: The case of SABIC price index. Res. J. Bus. Manage., 6: 121-133.

Boser, B.E., I.M. Guyon and V.N. Vapnik, 1992. A training algorithm for optimal margin classifiers. Proceedings of the 5th Annual Workshop on Computational Learning Theory, July 27-29, 1992, Pittsburgh, Pennsylvania, USA., pp: 144-152.

Cortes, C. and V. Vapnik, 1995. Support-vector networks. Mach. Learn., 20: 273-297.

Fama, E.F., 1965. The behavior of stock-market prices. J. Bus., 38: 34-105.

Hosmer, D.W. and S. Lemeshow, 2000. Applied Logistic Regression. 2nd Edn., Wiley and Sons, New York.

Hsu, C.W., C.C. Chang and C.J. Lin, 2003. A practical guide to support vector classification. Technical Report, University of National Taiwan, Department of Computer Science and Information Engineering, July, 2003. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf

Long, J.S., 1997. Regression Models for Categorical and Limited Dependent Variable. Sage Publications, Thousand Oaks, CA.

Ou, P. and H. Wang, 2009. Prediction of stock market index movement by ten data mining techniques. Mod. Appl. Sci., 3: 28-42.

Pohar, M., M. Blas and S. Turk, 2004. Comparison of logistic regression and linear discriminant analysis: A simulation study. Metodoloski Zvezki, 1: 143-161.

Ripley, B.D., 1996. Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge.

Suykens, J.A.K. and J. Vandewalle, 1999. Least squares support vector machine classifiers. Neural Process. Lett., 9: 293-300.

Venables, W.N. and B.D. Ripley, 2002. Modern Applied Statistics with S. 4th Edn., Springer-Verlag Inc., New York, ISBN: 0-387-95457-0.

Yuan, C., 2011. Predicting S&P 500 returns using support vector machines: Theory and empirics. October 12, 2011. http://apps.olin.wustl.edu/cres/research/calendar/files/YuanPaper.pdf