

Stock Feature Extraction Using Principal Component Analysis

Mbeledogu.N.N, Odoh.M.⁺ and Umeh.M.N

Department of Computer Science, Nnamdi Azikiwe University, Awka

Abstract. Reducing the dimensionality of feature vector is the most direct way to solve the problems caused by high feature dimensionalities. This is normally achieved in feature extraction step in data mining or pattern recognition system. The main task of feature extraction is to select or combine the features that preserve most of the information and remove the redundant components in order to improve the efficiency of the subsequent classifiers without degrading their performances. The Principal Component Analysis Technique was used to reduce 19 stock data variables to 9 stock data variable for stock prediction system. The result exhibited PCA's advantage of quantifying the importance of each dimension for describing the variability of a data set.

Keywords: Data Mining, Feature Extraction, Pattern recognition and Principal Component Analysis.\

1. Introduction

In recent years, database technology has advanced in stride. Vast amounts of data have been stored in the databases and business people have realized the wealth of information hidden in those data sets. Data mining which is a set of computer-assisted techniques designed to automatically mine large volumes of integrated data for new patterns associations, anomalies, statistically significant structures, hidden or unexpected information, or patterns [1]; [2]; then became the focus of attention as it promises to turn those raw data into valuable information that can be used to increase their profitability such as enabling these companies to determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics.

The capability to deal with voluminous data sets does not mean data mining requires huge amount of data as input. In fact, the quality of data to be mined is more important. Aside from being a good representative of the whole population, the data sets should contain the least amount of noise-errors that might affect mining results as shown in fig. 1.

Stock market is an important area of financial forecasting, which is of great interest to stock investors, stock traders and applied researchers. Data obtained from stock market is a time-series data which is often characterized with high dimensionality. It is usually not enough to look at each point in time sequentially; rather, one has to deal with sliding windows of a possibly multi-dimensional time series. By enlarging the context windows, this quickly produces very high dimensional vectors, thereby introducing the curse of dimensionality [3]. Main issues in developing a fully automated stock market prediction system are: feature extraction from the stock market data, feature selection for highest prediction accuracy, the dimensionality reduction of the selected feature set and the accuracy and robustness of the prediction system [4].

The pre-processing of the data is a time-consuming, but critical first step in the data mining process. It is often domain and application dependent. It encompasses de-noising, object identification, feature extraction and normalization. Fig. 1 shows the steps involved in scientific data mining as adapted from [2].

⁺ Corresponding author.

E-mail address:oguzuruodo@gmail.com

Feature extraction is a mechanism that computes numeric or symbolic information from the observation. It's main task is to select or combine the features that preserve most of the information and remove the redundant components in order to improve the efficiency of the subsequent classifiers without degrading their performances. It is the process of acquiring higher level information [5]. The dimensionality of the feature space may be reduced by the selection of subsets of good features. Feature extraction plays an important role in a sense of improving classification performance and reducing the computational complexity [6]. It also improves computational speed due to the fact that for less features, less parameters have to be estimated.

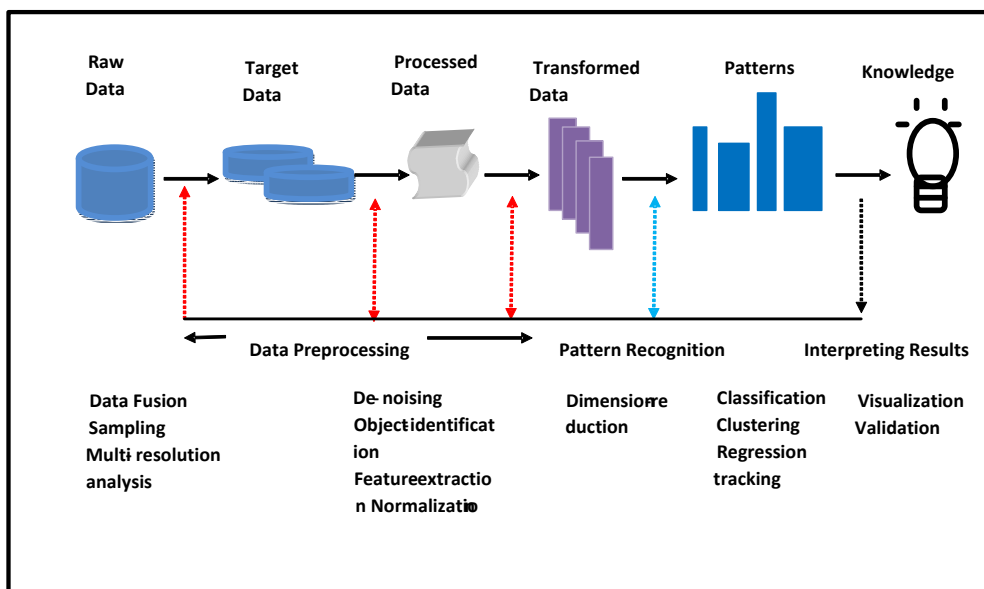


Fig. 1 Scientific Data mining

2. Background

Often in multivariate analysis, a fairly large number, p , of correlated random variables X are dealt with. It would be useful if it could be reduced to smaller number of random variables Y_1, Y_2, \dots, Y_s such that

1. The Y 's account for a large part of the total variability of the X 's.
2. The Y 's are interpretable in terms of the original problem.
3. The Y 's are independent.

The general objectives of principal component analysis are data reduction and data interpretation [7]; [8]; [9] considered Principal Component Analysis (PCA) as a popular independent feature extraction algorithm. It can reveal relationships that were not previously suspected thereby allowing interpretations that would not ordinarily result [7]. Principal Component Analysis (PCA) represents a powerful tool for analyzing data by reducing the number of dimensions, without important loss of information and has been applied on datasets in all scientific domains [10]. The purpose of PCA is to reduce the large dimensionality of the data space (observed variables) to the smaller intrinsic dimensionality of feature space (independent variables), which are needed to describe the data economically. Analyses of PCA more often than not serve as intermediate steps in much larger investigations such as inputs to a multiple regression or cluster analysis [7]. This is the case when there is a strong correlation between observed variables. The jobs which PCA can do are prediction, redundancy removal, feature extraction, data compression etc.

3. Review of Literature

Principal Component Analysis technique has been extensively used in many research works: [11] described the joint structure with a model that can potentially be used for scenario analysis and for estimating the risk of interest rate-sensitive portfolios. Three variations of the PCA technique to decompose global yield curve and interest rate implied volatility structure were examined and they concluded that global yield curve

structure can be described with 15–20 factors, whereas implied volatility structure requires at least 20 global factors [11] applied principal component analysis in loan granting. The result emphasized the utility of PCA in the banking sector to reduce the dimension of data, without much loss of information [12] performed a selection of optimal SNP-sets that capture intragenic genetic variation. Their findings suggested that PCA may be powerful tool for establishing an optimal SNP set that maximizes the amount of genetic variation captured for a candidate gene using a minimal number of SNPs. [13] analyzed the structure of RRab star light curves using PCA. They concluded that PCA is a very efficient way to describe many aspects of RRab light curves.

4. Methodology

Principal Component Analysis was applied. The instrument used to apply the PCA technique was Minitab [14]. PCA is the oldest technique in multivariate analysis and was first introduced by Pearson in 1901. It involves a mathematical procedure that transforms a number of (possibly) correlated variables into a (smaller) number of uncorrelated variables called *principal components*. The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. It is an eigenvector/value-based approach used in dimensionality reduction of the multivariate data. It is a way of identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in data of high dimension, where the luxury of graphical representation is not available, PCA is a powerful tool for analyzing data [8]. The other main advantage of PCA is that once these patterns are found in the data, the data is then compressed without much loss of information. It is widely used in most of the pattern recognition applications like face recognition, image compression, and for finding patterns in high dimensional data.

The mathematical equations of PCA: Consider a set of n observations on a vector of p variables organized in a matrix X ($n \times p$):

$$\{X_1, X_2, \dots, X_n\} \in R^p \quad (1)$$

The PCA method finds p artificial variables (principal components). Each principal component is a “linear combination of X matrix columns, in which the weights are elements of an eigenvector to the data covariance matrix or to the correlation matrix, provided the data are centered and standardized”. The principal components are uncorrelated.

The first principal component of the set by the linear transformation is:

$$Z_1 = a_1^T x_j = \sum_{i=1}^p a_{i1} x_{ij}, j = 1, \dots, n \quad (2)$$

In equation (2), the vectors a_1 and x_j are:

$$a_1 = (a_{11}, a_{21}, \dots, a_{p1}) \quad (3)$$

$$x_j = (x_{1j}, x_{2j}, \dots, x_{pj}) \quad (4)$$

One chooses a_1 and x_j such as the variance of z_1 is maximum. All principal components start at the origin of the ordinate axes. First PC is direction of maximum variance from origin, while subsequent PCs are orthogonal to first PC and describe maximum residual variance.

The main steps of PCA algorithm Fig. 2 (Ioniță and Șchiopu, 2010).

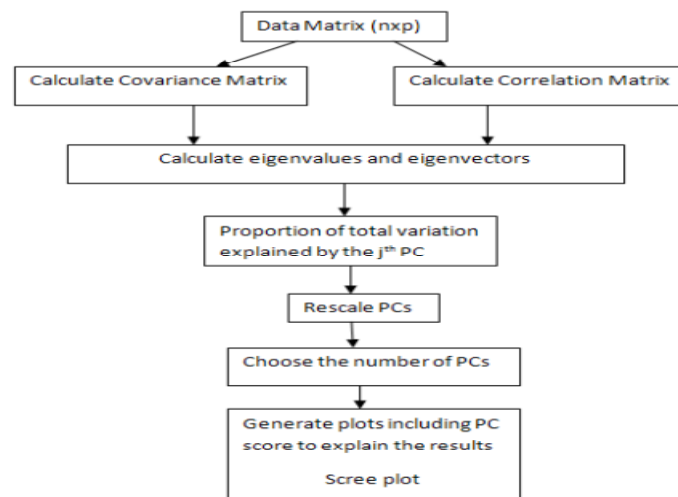


Fig. 2 Principal Component Analysis Steps

5. Specific Method

We obtained our daily stock dataset of 300 records from public database that contains Nigerian stock exchange data [15]. Table 1 presented the first five rows of the original stock dataset. The dataset contains the different variables- Day's price, A day's avg. price, A day's change in absolute terms, A day's change in %, marked down price at closure date, Day's high, Day's low, Year low, Year high, YTD change(%), EPS, Earnings yield at current price in %, P/E ratio, Final DPS, Interim DPS, Dividend yield at current price (%),Deals traded, Volumes traded and value traded for C1 – C19 respectively.

Table 1: original stock dataset

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19
40	39.6	1.05	2.7	36.1	40.1	38.5	38.95	40	2.7	2.16	5.41	18.49	1.15	0.4	2.88	139	3455312	136710552.1
40	40	0	0	36.1	40.2	40	38.95	40	2.7	2.16	5.41	18.49	1.15	0.4	2.88	212	5635864	0
41	40.4	0.99	2.48	36.1	41	39.5	38.95	40.99	5.44	2.16	5.28	18.94	1.15	0.4	2.81	204	4702993	189912896.9
41.9	41.2	1.01	2.47	36.1	41.9	40.7	38.95	41.89	7.55	2.16	5.17	19.36	1.15	0.4	2.75	299	6612358	272357208.9
43	41.7	1.09	2.6	36.1	43	40	38.95	42.99	10.37	2.16	5.03	19.87	1.15	0.4	2.68	310	12635911	526917118.8

6. Results and discussion

Columns with approximately constant values are ignored which reduced the data point from 19 to 14. The variables removed are: Marked down price@ closure date, Year low (N), EPS, Final DPS and Interim DPS. This is so as constant value along the column will make it impossible to perform the required analysis and prompt error message.

Table 2 contains the reduced data set. The variables are labeled C1, C2, ..., C14 which corresponds to Day's price, A day's avg. price, A day's change in absolute terms, A day's change in %, Day's high, Day's low, Year high, YTD change(%), Earnings yield at current price in %, P/E ratio, Dividend yield at current price (%),Deals traded, Volumes traded and value traded respectively.

Table 2 : Reduced stock data set

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14
40	39.6	1.05	2.7	40.1	38.5	40	2.7	5.41	18.49	2.88	139	3455312	136710552.1
40	40	0	0	40.2	40	40	2.7	5.41	18.49	2.88	212	5635864	0
41	40.4	0.99	2.48	41	39.5	41	5.44	5.28	18.94	2.81	204	4702993	189912896.9
41.9	41.2	1.01	2.47	41.91	40.7	41.9	7.55	5.17	19.36	2.75	299	6612358	272357208.9
43	41.7	1.09	2.6	43	40	43	10.37	5.03	19.87	2.68	310	12635911	526917118.8

Presented in Table 3 is the Eigen- analysis of the correlated matrix obtained from Table 2

Table 3 Eigen analysis of the Correlation Matrix

Eigenvalue: 8.3863 1.8397 1.1668 0.9710 0.8433 0.3834 0.2281 0.0819

Proportion: 0.599 0.131 0.083 0.069 0.060 0.027 0.016 0.006

Cumulative: 0.599 0.730 0.814 0.883 0.943 0.971 0.987 0.993

Eigenvalue: 0.0703 0.0228 0.0053 0.0008 0.0002 0.0000

Proportion: 0.005 0.002 0.000 0.000 0.000 0.000

Cumulative : 0.998 1.000 1.000 1.000 1.000 1.000

From the above Eigen analysis, it can be deduced that the values reduced from beginning to the last and significantly level up from 11 to 14 but the cumulative value of variable labeled 10 made it part of insignificant component as displayed on the screen plot.

Variable	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
C1	0.340	0.039	-0.080	-0.014	0.073	-0.092	0.114	-0.262
C2	0.327	0.050	-0.116	-0.026	0.115	0.073	0.163	0.771
C3	0.156	-0.634	0.192	-0.033	-0.107	0.016	0.067	0.046
C4	0.132	-0.656	0.202	-0.038	-0.109	-0.001	0.113	-0.017
C5	0.341	0.008	-0.014	0.001	0.020	-0.023	-0.056	0.149
C6	0.325	0.055	-0.151	-0.013	0.196	-0.014	-0.038	0.236
C7	0.292	-0.013	0.062	0.023	0.096	0.691	-0.609	-0.159
C8	0.340	0.040	-0.080	-0.014	0.073	-0.092	0.114	-0.263
C9	-0.338	-0.051	0.091	0.003	-0.086	0.156	-0.143	0.268
C10	0.092	-0.013	-0.531	-0.359	-0.747	-0.002	-0.143	0.012
C11	-0.338	-0.055	0.090	0.009	-0.083	0.147	-0.137	0.275
C12	0.169	0.327	0.450	0.158	-0.471	0.400	0.487	-0.018
C13	0.218	0.177	0.509	0.090	-0.275	-0.541	-0.504	0.117
C14	0.018	-0.125	-0.332	0.913	-0.187	-0.037	-0.059	0.015

Variable	PC9	PC10	PC11	PC12	PC13	PC14
C1	0.114	0.082	-0.506	-0.062	0.014	0.708
C2	0.474	0.106	0.009	-0.024	-0.002	-0.000
C3	-0.042	0.045	-0.087	0.711	0.016	0.002
C4	-0.008	0.050	0.113	-0.686	-0.011	-0.002
C5	-0.266	-0.885	-0.051	-0.031	-0.003	-0.000
C6	-0.777	0.407	0.042	-0.032	0.014	0.000
C7	0.139	0.062	0.061	-0.013	-0.011	-0.000
C8	0.113	0.082	-0.508	-0.059	0.020	-0.706
C9	-0.118	-0.022	-0.451	-0.093	0.724	0.004
C10	-0.010	0.040	0.014	-0.006	-0.003	-0.000
C11	-0.137	-0.008	-0.503	-0.073	-0.688	-0.003
C12	-0.128	0.044	0.020	0.007	-0.001	0.000
C13	0.066	0.118	-0.002	-0.016	0.001	-0.000
C14	0.014	0.011	-0.006	-0.006	0.004	-0.000

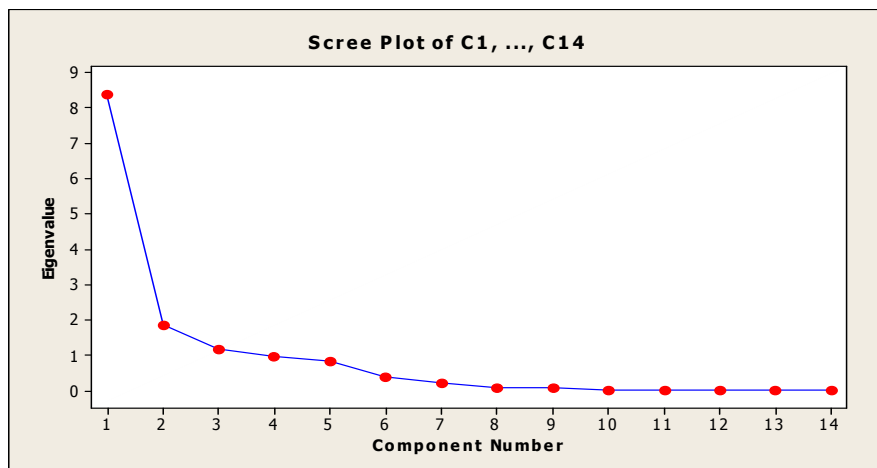


Fig. 3 Screen plot for identification of principal components

According to [7], one adhoc method is to use a screen plot. The screen plot is a plot of the eigenvalues along the y-axis against the number of principal components along the x-axis. To know the number of eigenvalues to keep, follow the graph until at a certain point, where it will level off, then, use the number of principal components up to that point. From fig.3, the plot leveled up at point 10 to 14 which limited the eigenvalues to be used to 9 instead of the selected 14. The corresponding variables selected are Day's price,

A day's avg. price, A day's change in absolute terms, A day's change in %, Day's high, Day's low, Year high, YTD change (%), Earnings yield at current price in % and P/E ratio.

7. References

- [1] Dai.J. Guha.R. and Lee.J.(2009). Efficient virus detection using dynamic instruction sequences. *Journal of computers*, vol.4, No. 5, 405-414. www.academypublisher.com/jcp/vol04/n.05/jcp0405405414.pdf
- [2] Kamath.C.(2005). Scientific data mining and pattern recognition: overview. Retrieved from <https://computation.llnl.gov/casc/sapphire/overview.html>
- [3] Mörchen Fabian(2003). Time series feature extraction for data mining using DWT and DFT. Retrieved from <http://mybytes.de/papers/moerchen03time.pdf>
- [4] Nair.B.B, Minuvarthini.M., Sujithra.B. and Mohandas.V.P(2010). Stock market prediction using a hybrid neuro-fuzzy system. *International Conference of Advances in Recent Technologies in Communication and Computing (ARTCom)*, Kottayam, 243-247. DOI:10.1109/ARTCom.2010.76
- [5] Alsutanny.Y.A and Aqel.M.M.(2003). Pattern recognition using multilayer neural-genetic algorithm. *Neurocomputing* 51,237-247. Retrieved from <http://Perso.telecom-paristech.fr/~bloch/p6/PRREC/abbas.pdf>
- [6] [Lei.X. (2003). A novel feature extraction method assembled with PCA and ICA for network intrusion detection. *Computer science- Technology and Applications IFCSTA* vol.3,31-34.
- [7] Onyeagu.S.I.(2003). *A first course in multivariate statistical analysis*. Mega Concept publishers, Awka, Nigeria. ISBN: 978-2272-29-9
- [8] Husnain.M. and Naweed.S.(2009). English letter classification using Bayesian decision theory and feature extraction using principal component analysis. *European journal of scientific research*. Retrieved from www.eurojournals.com/ejsr_34_2_06.pdf
- [9] Xuechuan.W.(2002). Feature extraction and dimensionality reduction in pattern recognition and their application in speech recognition. PhD dissertation. Retrieved from <http://www4.gu.edu.au:8080/adt-root/uploads/approved/adt-101front.pdf>
- [10] Ioniță.I. and Șchiopu.D.(2010). Using principal component analysis in loan granting. Retrieved from http://bmif.unde.ro/does/2010/pdf_final10IIonita-final2.pdf
- [11] Novosyolov.A. and Satchkov.D.(2008). Global term structure modeling using principal component analysis. *Journal of Asset Management*, Vol.9,1,49-60. Retrieved from <http://risktheory.ru/papers/jam20083a.pdf>
- [12] Horne.B.D. and Camp.N.J (2004). Principal component analysis for selection of optimal SNP-Sets that capture intragenic Genetic variation. *Genetic Epidemiology* 26: 11-21. DOI:10.1002/gepi.10292
- [13] Kanbur.S.M and Mariani.H.(2004). Principal component analysis of RR lyrae light curves. Retrieved from www.astro.umass.edu/~shashi/paper7.pdf
- [14] Hoffman.L. and Joseph.M.(2003). A multivariate statistical analysis of the NBA. Retrieved from www.units.muohio.edu/sumsri/sumj/2003/NBAstats.pdf
- [15] FSDH(2008). First Security Discount House (FSDH) securities limited. Retrieved from www.fsdhsecurities.com