Clustering underlying stock trends via non-negative matrix factorization

Andrea Pazienza¹, Sabrina Francesca Pellegrino², Stefano Ferilli¹, and Floriana Esposito¹

 ¹ Dipartimento di Informatica – Università di Bari
 ² Dipartimento di Matematica – Università di Bari {name.surname}@uniba.it

Abstract. Building a diversified portfolio is an appealing strategy in the analysis of stock market dynamics. It aims at reducing risk in market capital investments. Grouping stocks by similar latent trend can be cast into a clustering problem. The classical K-Means clustering algorithm does not fit the task of financial data analysis. Hence, we investigate Non-negative Matrix Factorization (NMF) techniques which, contrary to K-Means, turn out to be very effective when applied to stock data. In particular, recently developed NMF techniques, which incorporate convexity constraints, generate more disjoint latent trend groupings than the traditional sector-based groupings. In this paper, the NMF technique and its variants are applied to NASDAQ stock data (i.e., daily closing prices). Experimental results confirm that (*convex*) NMF techniques are highly recommended to produce trend based assets and build a good diversified portfolio.

1 Introduction

A trader's purpose is to beat the market and, then, to make money. To achieve this objective, the trader should be able to predict future stock prices. In this way, he could determine a self-financing trading strategy that maximizes his portfolio return [5]. However, because of randomness in the market, creating and managing successful portfolios of financial assets is a difficult practice. Diversification theory is the most widely used practice by individuals to develop portfolios. It is based on the principle of attempting to maximize expected return for a given amount of risk or, equivalently, to minimize the risk for a given amount of return [9]. It is one of the most effective ways to get low risk-reward ratios. This problem can be seen as a clustering process, in which the aim is grouping data (e.g., stocks) into subgroups of similar behavior (e.g., market trend).

Clustering is arguably one of the most common steps in unsupervised behavior analysis. With K-Means [8], a classical clustering algorithm, it is not possible to establish the effectiveness and coherence of the clusters when dealing with stock data. Therefore, more powerful analysis techniques are required. Matrix decomposition strategies can overcome this problem, in fact, they can provide ways to produce cleaner data which may lead to a better interpretation of the results. Since the closing prices of stocks are definitely non-negative signals, it makes sense to apply Non-negative Matrix Factorization (NMF) on them.

In this paper, we use NMF and its variants to learn the components which drive the stock market, and to construct a diversification method using cluster analysis of financial assets. We compare NMF with its two variants, Convex NMF (C-NMF) and Convex-Hull NMF (CH-NMF), obtained by imposing orthogonal and convex constraints. We investigate the impact of these constraints in real stock return data.

Finally, we conclude that CH-NMF yields a more accurate and disjoint representation of the data, and allows a better interpretation of clustering. Moreover, CH-NMF is very useful because it converges in only one iteration, with very low runtime and, most important, achieving a very small error in Frobenius norm.

This paper is organized as follows. The next section recalls useful background information, including related works. Then, Section 3 introduces the mathematical aspects of NMF, its variants C-NMF and CH-NMF. Section 4 describes experimental results and evaluates the difference between the methods in terms of number of iterations, error and clustering representation. Finally, Section 5 concludes the paper.

2 Background and Related Work

Portfolio Diversification is the process of choosing investments in order to reduce exposure to a particular asset. This is typically done by investing in a variety of assets, because, if the stock prices do not move together, then a diversified portfolio of assets will have lower variance than the weighted average variance of the assets. A straightforward diversification breaks the stock market into several classic sectors according to the primary activities of a company (such as Basic Materials, Technology, Financial, Health Care, etc.).

As the stock evolution is comparable to a stochastic process, stock prices are determined by fluctuations in underlying or latent trends, which can be modeled by a Brownian motion [5]. Therefore, stocks in the same sector may not show similar behavior in the market. So, albeit the sector-based strategy is simple enough to apply, the portfolio will not ensure the maximum return. Hence, if one were able to identify and predict the underlying trends from the stock market data, one would have the opportunity to leverage this knowledge to obtain genuine portfolio diversification opportunities. In other words, investors should diversify their money into not only different sectors, but also different trends (e.g. *clusters*).

Unfortunately, the K-Means clustering algorithm, has still some limitations in the exploitation of financial data. Indeed, in [1] is stated that K-Means clustering tends to find spherical clusters so that centroid-based clustering does not handle the noise. Hence, the authors aimed to discover other centroid-based clustering approaches for financial datasets and to introduce weighted Euclidean distance instead of standard Euclidean distance to re-evaluate centroid-based clusters, as to overcome the limitations of K-Means. Matrix decompositions, especially NMF, are used in the literature for the analysis of financial data. In particular, [4] applied a constrained NMF to real stock data and found that there is a tradeoff between smoothness of trend and sparsity of the weight matrix. [2] provides a contribution to the Diversification Theory by comparing Semi-NMF with Sparse-semiNMF applied to a diffusive model based on the Black&Scholes equation for option pricing. It is deduced that Sparse-semiNMF outperforms semi-NMF because it better reduces the risk related to the portfolio selection. Using the multiplicative update rules algorithm, [7] analyzes the behavior of latent trends for different value of the number of the latent forces and shows that the increase in the underlying force does not affect the trends of the original forces. In [11] a variant of Sparse-semiNMF with sum-to-one and smoothness constraints is applied to the portfolio diversification problem, and results show a disjoint data representation that allows a better understanding of stock properties.

In our setting, assume the market is made up of m stocks S_1, S_2, \ldots, S_m ; each stock S_i is stored as a row vector whose entries are n daily closing prices. Suppose there are k latent bases, W_1, W_2, \ldots, W_k ; each W_j is a n-dimensional row vector, which can be thought as a Brownian motion. So it is possible to express each stock as linear combination of these bases:

(1)
$$S_i = \sum_{j=1}^k h_{ij} W_j$$

where h_{ij} is a non negative real number and indicates the association degree of the *i*-th stock with the basis W_j . With matrix notation, (1) can be expressed as

$$S_{+} = H_{+} W_{\pm}$$

where $S \in \mathbb{R}^{m \times n}_+$, $H \in \mathbb{R}^{m \times k}_+$ (weight matrix) and $W \in \mathbb{R}^{k \times n}_{\pm}$ (trend matrix). This strongly recalls the non-negative matrix factorization formulation.

3 Non-Negative Matrix Factorization

The standard definition for non-negative matrix factorization of a matrix S is

$$(3) S = H W,$$

where $S \in \mathbb{R}^{m \times n}$, $H \in \mathbb{R}^{m \times k}$ and $W \in \mathbb{R}^{k \times n}$, and $k \leq m$. Both W and H must contain only non-negative entries. W is the matrix of factors and H is the mixing matrix.

According to [6], each data point, which is represented as a row in S, can be approximated by an additive combination of the non-negative basis vectors, which are represented as rows in W, weighted by the components of H. Matrices H and W are found by solving the optimization problem

(4)
$$\min_{\substack{H \ge 0, \\ W \ge 0}} \|S - H W\|_F^2,$$

where $\|\cdot\|_F$ is the Frobenius norm. The algorithm is expressed in terms of a pair of update rules that are applied alternately:

$$H_{ij} = H_{ij} \sum_{k} \frac{S_{ik}}{(HW)_{ik}} W_{jk}, \qquad W_{ij} = W_{ij} \sum_{k} H_{ki} \frac{S_{kj}}{(HW)_{kj}}$$

Matrices H and W are initialized at random. Various variants and improvements to NMF have been introduced in recent years [3, 10].

Convex NMF Convex non-negative matrix factorization (C-NMF) [3] allows the data matrix S to have mixed signs. It minimizes $||S - SHW||_F^2$ subject to the convex constraint $||H_i||_1 = 1$, $H \ge 0$, where $S \in \mathbb{R}^{m \times n}$, $H \in \mathbb{R}^{n \times k}$ and $W \in \mathbb{R}^{k \times n}$. Matrices H and W are updated iteratively, until convergence, using the following update rules:

$$H_{ik} = H_{ik} \sqrt{\frac{(Y^+ W)_{ik} + (Y^- H W^T W)_{ik}}{(Y^- W)_{ik} + (Y^+ H W^T W)_{ik}}}$$
$$W_{ik} = W_{ik} \sqrt{\frac{(Y^+ H)_{ik} + (W H^T Y^- H)_{ik}}{(Y^- H)_{ik} + (W H^T Y^+ H)_{ik}}}$$

where $Y = S^T S$, and matrices Y^+ and Y^- are given by $Y_{ik}^+ = \frac{1}{2} (|Y_{ik}| + Y_{ik})$ and $Y_{ik}^- = \frac{1}{2} (|Y_{ik}| - Y_{ik})$ respectively. Matrices H and W are initialized at random. The convex constraint imposed on H has the advantage that one might interpret the rows of H as weighted sums of certain data points. This means that these rows can be interpreted as centroids. Moreover, C-NMF solutions are gen-

erally significantly more orthogonal than Semi-NMF solutions.

Convex-Hull NMF Massive datasets are likely to capture even very rare aspects of the problem at hand. Along this line, [10] recently introduced a datadriven Convex NMF approach, called Convex-Hull NMF (CH-NMF), that is fast and scales extremely well: it can efficiently factorize huge matrices and in turn extract meaningful "clusters" from massive datasets. The key idea is to restrict the clusters to be combinations of vertices of the convex hull of the dataset; this allows to directly explore the data itself to solve the convex NMF problem.

We consider a factorization of the form S = S H W where $S \in \mathbb{R}^{m \times n}$, $H \in \mathbb{R}^{n \times k}$ and $W \in \mathbb{R}^{k \times n}$. We further restrict the rows of H and W to convexity, i.e., $||H_i||_1 = 1$ $H \ge 0$, $||W_j||_1 = 1$ $W \ge 0$.

In contrast to C-NMF, we consider a convex combination on both H and W. The task now is to minimize

(5)
$$||S - SHW||_F^2$$
, s.t. $||H_i||_1 = 1$, $H \ge 0$, $||W_j||_1 = 1$, $W \ge 0$.

This optimization problem is equivalent to projecting the solution in the convex hull of S. Convexity constraints yield latent components with some properties: first, any data point can be expressed as a convex and meaningful combination of these basis vectors; this offers interesting new opportunities for data

interpretation. Second, they span a simplex that encloses most of the remaining data. CH-NMF aims at a data factorization based on the data points residing on the data convex hull. Therefore, CH-NMF seeks an approximate solution by subsampling the convex hull, exploiting each data point on the convex hull as a linear lower dimensional projection of the original data.

A consequence of the convexity of H and W is that the rows of H tend to become nearly orthogonal. Requiring orthogonal rows for H produces a noncorrelation between stocks being attracted from different clusters. This indicates a more accurate clustering and, hence, that the NMF family is competitive with K-Means for the purposes of clustering financial data. Therefore, the aim of this paper is to exploit NMF techniques, and in particular the ones with convexity constraints, in the context of financial data.

4 Evaluations and Comparisons of NMF Techniques

The objective of this work is studying the application of clustering approaches to determine a trend-based portfolio diversification that is more consistent than the sector-based one. We ran experiments on stock data gathered from the NASDAQ Stock Market (http://www.nasdaq.com), and specifically on the closing prices of 28 stocks in the past 10 years (2518 working days). Table 1 reports the list of stocks involved in the experiments and their belonging sector. Fig. 1(a) shows the actual stock prices, useful as a reference to compare the numerical solutions that we obtain from the different methods. So, we want to identify the subgroups (clusters) of stocks that show a similar trend. In particular, we want to assess the performance of NMF, C-NMF and CH-NMF. We try different numbers k of clusters. Since the considered stocks involve only 8 market sectors overall, we ran the methods 6 times, one for each $k \in \{3, 4, \ldots, 8\}$. We used a Python implementation of these methods available at http://pymf.googlecode.com.

The raw data were preprocessed as *log-returns* and stored into a matrix $S \in \mathbb{R}^{28 \times 2518}$. In this way the variances of stock data distribution are homogeneous to allow a better understanding of the graphical results. We compute two matrices H and W for the given stock data matrix S. H and W are used to identify the k cluster labels of the stocks: in fact, rows of W regard the cluster centroids while H is the cluster membership indicator matrix. In other words, sample i is in cluster j if H_{ij} is the largest value in row $H_{i,:}$. The product HW, representing the reconstruction of S, is a useful way to explicit the difference between the original stock prices data and the approximated dynamics of transformed data.

Table 2 reports, for each k, statistics about the results obtained by the NMF methods: number of iterations for convergence, error estimation in Frobenius norm, and number of non-empty clusters in the result. We note that, in most cases, all decomposition methods provide a good reconstruction of data. They differ in the cost to achieve a good decomposition (the larger the number of iterations, the more expensive the method). We want to highlight that CH-NMF is very fast and its estimated Frobenius error is of the same order of magnitude as the other methods. In fact, even if we set a higher number of maximum iterations,

Table 1: Stocks Table

#	Code	Company	Sector
1	AA	Alcoa Inc.	Basic Materials
2	AIG	American International Group, Inc.	Financial
3	AAPL	Apple Inc.	Technology
4	AXP	American Express Company	Financial
5	BA	Boeing Company	Industrial Goods
6	CAT	Caterpillar, Inc.	Industrial Goods
7	DD	E.I. du Pont de Nemours and Company	Basic Materials
8	DIS	Walt Disney Company	Services
9	GE	General Electrics Company	Conglomerates
10	HD	Home Depot, Inc.	Services
11	HON	Honeywell International Inc.	Industrial Goods
12	HPQ	HP Inc.	Technology
13	IBM	International Business Machines Corporation	Technology
14	INTC	Intel Corporation	Technology
15	JNJ	Johnson & Johnson	Healthcare
16	JPM	JP Morgan Chase & Co.	Financial
17	KO	Coca-Cola Company	Consumer Goods
18	MCD	McDonald's Corporation	Services
19	MSFT	Microsoft Corporation	Technology
20	MMM	3M Company	Conglomerates
21	MO	Altria Group, Inc.	Consumer Goods
22	PFE	Pfizer, Inc.	Healthcare
23	PG	Procter & Gamble Company	Consumer Goods
24	UTX	United Technologies Corporation	Conglomerates
25	VZ	Verizon Communications Inc.	Technology
26	WMT	Wal-Mart Stores, Inc.	Services
27	ABIO	ARCA biopharma, Inc.	Healthcare
28	AMGN	Amgen Inc.	Healthcare

 Table 2: Numerical results

		NMF			C-NMF			CH-NMF		
	k	# iter	error	# clusters	# iter	error	# clusters	# iter	error	# clusters
ĺ	3	1528	33.7703	3	500	45.7185	2	1	47.5844	2
	4	2355	27.1966	4	500	42.4148	2	1	43.9824	4
	5	3358	21.0838	5	500	40.2502	2	1	38.6585	4
	6	2523	16.9987	4	500	33.5761	2	1	56.4050	4
	7	5000	14.6706	6	500	38.6675	2	1	32.5755	5
	8	5000	13.5482	7	500	32.1786	2	1	46.2535	5



Fig. 1: Original stock data with reconstruction for k = 4

CH-NMF needs only 1 iteration to converge for these particular dataset, as reported in Table 2. For others techniques involved in the experiments, NMF requires up to 5000 iterations in the worst case, while C-NMF takes less iterations than NMF to converge, but it always discovers only two non-empty clusters. This is very unattractive because, considering only 2 trends will expose investors to high risks as they would diversify their portfolio making it up with only 2 stocks, one for each trend. The role of k is to force a representation for the data that is more compact than its actual form. Assuming a more compact representation will capture underlying regularities in the data that might be obscured by the form in which the data is found in matrix S. The target is to achieve a low rank approximation which ensures a good interpretation of data in terms of clustering partition, data reconstruction and compactness of representation. Indeed, for a genuine portfolio diversification, choosing k = 4 represents a fair compromise between a lower rank approximation and the goal of yielding a good cluster partition. The reconstruction obtained for k = 4 is shown Fig. 1(b,c,d).

Now, we focus on the study of W, i.e., the matrix which represents the latent trends. As shown in Fig. 2, the trend obtained from NMF points out the increment of fluctuations as k grows up, despite the quality of graphical reconstruction being good. In Fig. 3 we can see that C-NMF shows too many fluctuations and does not allow us to compose a good diversified portfolio. Indeed, as shown in Table 2, for each k, C-NMF provides always only 2 clusters, which reflect the trends that can be seen in the figures. A reason for this behavior is related to the fact that C-NMF imposes the convexity constraint only on H. Hence, convexity for H should be used together with convexity also on W. In fact, CH-NMF is able to overcome this problem and leads to more regular components (cfr. Fig. 4).







Fig. 3: Trends for C-NMF



Fig. 4: Trends for CH-NMF

Therefore, we can state that the convexity constraint of H and W provides a good adjustment: the bases become more disjoint and Frobenius norm decreases at a speed that depends on the number of iterations. It is important to note that while all methods try to minimize the same criterion, they impose different constraints and thus yield different matrix factors. For example, CH-NMF assumes W and H to be non-negative matrices and often leads to sparse representations of the data.

Another important graphical confirmation of our proposal can be found in the analysis of colormaps which is a good way to display matrices in scaled colors. It represents a color-filled table in which every color indicates the weight of each corresponding matrix component according to a total order relation managed by a color scale called colorbar. In this way, we can evaluate the components of greater weight associated with latent trends. More precisely, every H_{ij} indicates how the *i*-th stock is related to the row basis W_j . Also in this representation, we can observe the degree of belonging of each data point to a cluster by selecting, for each row, the highest element in the colormap. In the case of NMF (see



Fig. 5: Colormaps for k = 4

Fig. 5(a)), colormaps display a regular proceeding of data with a high peak in some rows. This determines the membership of an element in a row to a cluster in the corresponding column. While, for C-NMF (see Fig. 5(b)) the elements with the highest color with respect to the colorbar are located in the first or last column, giving more emphasis to the fact that the resulting clusters are always 2. Regarding colormaps for CH-NMF (see Fig. 5(c)), we can observe a clearer disjunction of columns, meaning that the resulting clusters are readily visible.

Table 3: NMF Clusters

k = 3	k=4 $k=5$		k = 6	k = 7	k = 8			
$3\ 4\ 5\ 6$	$3\ 5\ 8\ 10$	$12 \ 14 \ 15$	3456	$3\ 4\ 5\ 6$	11 15 17 18			
7 8 9 10	11 15 16	$16\ 17\ 18$	7 8 10 11	7 8 14 16	20 <mark>21</mark> 22 <mark>23</mark>			
11 14 15 16	19 20 21	19 20 <mark>23</mark>	13 14 15 16	$19\ 20\ 24$	$24 \ 25 \ 26 \ 28$			
17 19 20 <mark>21</mark>	22 23 24	24 25 <mark>26</mark>	17 18 19 20					
22 23 24 25	25 26 28		21 22 24 25					
26 28			28					
12 13 18	671213	$4\ 5\ 8\ 10$	23 26	12 15 23	$2\ 4\ 5\ 8$			
	14 17 18	$21 \ 22 \ 28$		26	$9\ 10$			
1 2 27	27	7 11	1 2 9 27	21 28	27			
	1249	3 6 13	12	10 13 17	16			
				$18 \ 22 \ 25$				
		$1 \ 2 \ 9 \ 27$		27	3 13 14			
				$1 \ 2 \ 9 \ 11$	$1\ 6\ 7\ 12$			
					19			

Compared to the K-Means algorithm, used as baseline, the main difference lies into the creation of clusters. In fact, K-Means always produces exactly kclusters, while NMF methods generate *at most* k clusters, as shown in Table 2. This means that there are centroids which are not attracting stocks.

After analyzing all decompositions, our main purpose is to obtain clusters of stocks with the same trend, starting from a matrix decomposition of data in a such way that W would hold centroids coordinates and H would hold the relationship degree of different centroids. We collect every *i*-th row of H, which corresponds to the *i*-th stock index for Table 1, into their own membership cluster in order to discover, for each k, which subgroups of stocks remain unaffected. Thus, the most frequent subgroups of stock data can be chosen as final outcome of our portfolio diversification strategy. More precisely, it could be possible to construct a tempting portfolio by selecting the most promising stocks from each subgroup. We implemented this functionality in MATLAB with the objective to take out a cluster matrix by varying both of k and the decomposition method. In Tables 3-5, we show the resulting clusters of stock data: we see that clusters of stocks persist across the different decomposition methods. To give some concrete examples, stocks with index 1, 2, 9, 27 (and often 12 too) are always grouped together. They correspond to red indexes in the tables.

Table 4: C-NMF Clusters

k = 3	k = 4	k = 5	k = 6	k = 7	k = 8
1 2 12 27	$1 \ 2 \ 9 \ 12$	$1 \ 2 \ 9 \ 12$	$1\ 2\ 27$	1269	$1\ 2\ 9\ 12\ 27$
	27	27		12 23 27	
34567	$3\ 4\ 5\ 6\ 7$	$3\ 4\ 5\ 6\ 7$	$3\ 4\ 5\ 6\ 7$	$3\ 4\ 5\ 7\ 8$	$3\ 4\ 5\ 6\ 7$
8 9 10 11	8 10 11	8 10 11	8 9 10 11	10 11 13	8 10 11
13 14 15	$13 \ 14 \ 15$	$13 \ 14 \ 15$	$12 \ 13 \ 14 \ 15$	$14 \ 15 \ 16$	$13\ 14\ 15$
16 17 18	$16\ 17\ 18$	$16\ 17\ 18$	$16\ 17\ 18\ 19$	$17\ 18\ 19$	$16\ 17\ 18$
19 20 21	$19\ 20\ 21$	19 20 21	$20\ 21\ 22\ 23$	$20\ 21\ 22$	$19\ 20\ 21$
22 23 24	$22 \ 23 \ 24$	$22 \ 23 \ 24$	$24 \ 25 \ 26 \ 28$	$24 \ 25 \ 26$	$22 \ 23 \ 24$
25 26 28	$25 \ 26 \ 28$	$25 \ 26 \ 28$		28	$25 \ 26 \ 28$

They correspond respectively to Alcoa Inc. (Basic Materials), American International Group Inc. (Financial), General Electrics Company (Conglomerates), ARCA biopharma Inc. (Healthcare), HP Inc. (Technology). Other stocks which are often grouped together are depicted in the Tables with different colours.

Summing up, the obtained clusters demonstrate successful application of CH-NMF to the analysis of financial data. This means that CH-NMF is robust in the case of analysis of stock market. Moreover, it provides a trend-based diversification containing groups of different sectors. The most interesting result is that the stocks of the same sector is not necessarily assigned into the same cluster and vice versa, which is of potential use to guide diversified portfolio.

k=3 $k=4$		k = 5	k = 6	k = 7	k = 8			
1 2 9 12 27	1 2 9 27	1 2 9 27	$1 \ 2 \ 9 \ 12 \ 27$	129	2 27			
3 4 5 6 7 8	$3\ 4\ 5\ 7\ 8$	$3\ 5\ 7\ 8\ 10$	$3\ 4\ 5\ 7\ 8$	3467	4 8 10 11			
9 10 11 13 14	10 11 14 15	$11 \ 18 \ 21$	$11 \ 14 \ 15 \ 16$	13 14 17	$13\ 15\ 16\ 17$			
15 16 17 18 19	$16\ 17\ 18\ 19$	22 <mark>24</mark> 25	17 20 22 23	20 23 24	18 20 21 22			
20 21 22 23 24	20 21 22 23		24 25 <mark>26</mark> 28	25 26	23 24 25 26			
25 26 28	24 25 26 28				28			
	6 13	4 6 12 14	6 13	5 8 10 11	$1\ 5\ 9\ 12$			
		$15 \ 16 \ 17 \ 19$		15 16 18	14 19			
		20 23 26 28		19 21 22 28				
	12	13	10 18 19 21	12	67			
				27	3			

Table 5: CH-NMF Clusters

5 Conclusions

Constructing a diversified portfolio, in which the correlation between constituent asset classes and investment strategies is meaningfully low, can be challenging, in order to reduce the exposure to risk by investing in a variety of assets. Our aim is to group stocks having similar trend. This can be cast as a clustering problem in data mining that we solve with NMF techniques. We investigate NMF and its variants with convexity constraints to improve the exploitation of similar stock trends. In particular, we show that, for this task, CH-NMF is a very fast and scalable in terms of speed and reconstruction quality. Our extensive experimental evaluation shows that NMFs better point out the clustering properties, additionally yielding very low error in Frobenius norm and high efficiency in terms of convergence time. Furthermore, we compared the resulting clusters to check whether frequent itemsets of stock stay together still while the number of requested clusters changes. The task of prediction is not applicable for NMF techniques because the number of clusters to be discovered is given in input.

As future work, we will use more datasets from different markets and will investigate further decomposition techniques that can further improve the effectiveness of clustering stock data and impose other penalty constraints in order to achieve a better portfolio diversification strategy, reduce the risk of investments and, hence, beat the market.

Acknowledgments

This work was partially funded by the Italian PON 2007-2013 project PON02 00563 3489339 'Puglia@Service'.

References

- F. Cai, N. Le-Khac, and M. Kechadi. Clustering approaches for financial data analysis: a survey. In Proceedings of the International Conference on Data Mining (DMIN), pages 1–7, 2012.
- [2] R. De Fréin, K. Drakakis, and S. Rickard. Portfolio diversification using subspace factorizations. In *Information Sciences and Systems*, 2008. CISS 2008. 42nd Annual Conference on, pages 1075–1080. IEEE, 2008.
- [3] C. H. Q. Ding, T. Li, and M. I. Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, Jan 2010.
- [4] K. Drakakis, S. Rickard, R. De Fréin, and A. Cichocki. Analysis of financial data using non-negative matrix factorization. *International Mathematical Forum*, 3(38):1853–1870, 2008.
- [5] R. Korn and E. Korn. Option pricing and portfolio optimization. Graduate Studies in Mathematics, 31:18, 2001.
- [6] D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In Advances in neural information processing systems, pages 556–562, 2001.
- [7] T. Liu. Non-negative matrix factorization for stock market pricing. In *Biomedical Engineering and Informatics. BMEI'09.*, pages 1–5. IEEE, 2009.
- [8] J. MacQueen et al. Some methods for classification and analysis of multivariate observations. In Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, volume 1, pages 281–297. Oakland, CA, USA., 1967.
- [9] H. Markowitz. Portfolio selection. The journal of finance, 7(1):77–91, 1952.
- [10] C. Thurau, K. Kersting, and C. Bauckhage. Yes we can: simplex volume maximization for descriptive web-scale matrix factorization. In *Information and knowledge management. CIKM 2010.*, pages 1785–1788. ACM, 2010.
- [11] J. Wang. Stock trend extraction via matrix factorization. In Advanced Data Mining and Applications, pages 516–526. Springer, 2012.