Proceedings of the 5th Symposium on Physics in Economics and Social Sciences, Warszawa, Poland, November 25–27, 2010

Analysis of Time Series Correlation. The Choice of Distance Metrics and Network Structure

J. MIŚKIEWICZ

University of Wrocław, Institute of Theoretical Physics, pl. M. Borna 9, 50-204 Wrocław

Standard analysis of correlations between companies consists of two stages: calculating the distance matrix and construction of a chosen graph structure. In the paper the most often used Ultrametric Distance (UD) is compared with the Manhattan Distance (MD). It is showed that MD allows to investigate a broader class of correlation and is more robust to the noise influence. Therefore MD was used to construct entropy distance, which is applied to the analysis of correlation between subset of WIG20 and S&P500 companies. In the analysis three network structures were used: minimum spanning tree and unidirectional and bidirectional minimum length path. The results are compared to the standard UD based analysis. The advantages and disadvantages of the analysed time series distances are outlined.

PACS: 89.65.Gh, 05.45.Tp, 06.30.Bp

1. Introduction

The questions on the form of correlations between time series are important problems in various sciences. They are frequently raised in meteorology [1], medicine [2] or in economy [3, 4, 5]. One of the very basic features of economy systems is that the firms compete or cooperate and often they form hidden trusts. Such knowledge may be crucial when investing on a market, where massive losses due to a sudden fall of prices in a chosen group of shares make correlation investigations one of the most serious research topics [6]. Of course the mutual agreements are usually hidden and the only possible way to discover them is the analysis of the available data such as the officially announced information or the stocks market quotes. Besides the micro economy questions the correlation analysis might be also useful while rising macro economy issues e.g. measurement of the globalization process [7, 8]. For years the most popular methods for investigations of correlations is the Ultrametric Distance (UD) introduced by Mantegna [9] as an extension of the portfolio analysis [10] with various applications e.g. [11– 14]. The main disadvantage of the UD is that it measure only whether the linear correlation exits. It does not verify hypothesis of the correlation in general. On the other hand there are no reasons to constrain the correlation analysis only to the linear case.

Analytical form of the correlations is the not the only problem in this subject. An other question pertains to the methodology itself. The UD is very useful but only in the case where linear correlations are expected. Besides UD is sensitive to the presence of noise (as shown in Sec.3). But in contrast to classical measurement problems, e.g. in meteorology or economy, the noise is an intrinsic part of the data and the truly existing correlations may be hidden by the noise. Therefore a correlation measure which is not sensitive to noise is necessary to be found and used.

Within this paper two correlation measures are investigated and compared (Sec. 2): Ultrametric Distance and Manhattan Distance. Moreover Manhattan Distance is used to construct Entropy distance, the latter in the form of the Theil index based Manhattan Distance. The noise influence on the distance function is investigated in Sec.3, while the examples of the distances applications are presented in Sec. 4.

2. Distance measures

2.1. Ultrametric distance

The Ultrametric Distance $UD(A, B)_{(t,T)}$ is based on the Pearson linear correlation coefficient:

$$corr_{(t,T)}(A,B) =$$

$$\frac{\langle AB \rangle_{(t,T)} - \langle A \rangle_{(t,T)} \langle B \rangle_{(t,T)}}{\sqrt{(\langle A^2 \rangle_{(t,T)} - \langle A \rangle_{(t,T)}^2)(\langle B^2 \rangle_{(t,T)} - \langle B \rangle_{(t,T)}^2)}}, \qquad (1)$$

where the brackets $\langle \dots \rangle$ denote a mean value over a time window (t - T, t).

The correlation function Eq.(1) is transformed in order to fulfil the distance axioms:

$$UD(A,B)_{(t,T)} = \sqrt{\frac{1}{2}(1 - corr_{(t,T)}(A,B))}.$$
 (2)

The correlation function Eq.(1) takes the values in the interval [-1, 1]. The extremes values correspond to the anticorrelated and linearly correlated time series. The case $corr_{(t,T)}(A, B) = 0$ is interpreted as not correlated data series. However, it should be said in the latter case that the linear correlation was not found. Eq.(2) maps the linear space of the series L_n of length n onto the interval [0,1] $(UD(A, B) : L_n \times L_n \to [0,1])$. The

important points are $UD(A, B) = 0 \rightarrow$ linear correlation, $UD(A, B) = \frac{\sqrt{2}}{2} \approx 0.7 \rightarrow$ no correlation and $UD(A, B) = 1 \rightarrow$ anticorrelated time series. (The original formulation of Mantegna [9] was modified to obtain mapping on to the interval [0, 1] [15].)

2.2. Manhattan distance

The Manhattan distance known also as a city block distance or taxicab metric [16] received recently more attention in econophysics investigations [17, 18, 19]. For the sake of clarity the definition is recalled here. Let consider the time series A, B as vectors than the Manhattan Distance is defined by Eq.(3) as a sum of the absolute value of the difference between appropriate coordinates:

$$MD(A,B) = \sum_{i=1}^{n} |a_i - b_i|$$
(3)

where $A = (a_1, a_2, \dots, a_n), B = (b_1, b_2, \dots, b_n).$

Value of MD depend on the time series length. However in some applications of time series analysis one has to compare distances between time series in different time windows sizes. Therefore the mean Manhattan distance (MMD) is introduced here:

$$MMD(A,B) = \frac{1}{n} \sum_{i=1}^{n} |a_i - b_i|$$
(4)

The MD and MMD can be extended into an entropy distance measure by prior transformation of the time series into entropy time series and thereafter application of MD or MMD. However the entropy estimation in the case of economy time series is a difficult task. The most popular entropy formulation i.e. the Shannon entropy [20] requires knowledge of the probability distribution function. However, to estimate any PDF one needs a relatively large number of data. On the other hand taking time window of such a size one has to face two problems: one is that the length of the analysed time series is shortened significantly, another problem is the stationarity of economy time series. Most of the time series are nonstationary therefore it might be not possible to find with reasonable accuracy, the PDF of the given time series. The difficulties with PDF estimation can be work out by application of an alternative entropy definition. Within this paper the Theil index [21] defined by Eq.(5) is used.

$$Th_A(t,T) = \sum_{i=t-T}^{t} \left(\frac{a_i}{\sum_{j=t-T}^{t} a_j} \ln \frac{a_i}{\langle A \rangle_{(t,T)}}\right).$$
(5)

T is the time window length. The time series A transformed by the Theil index with the time window T will be denoted as Th(A, T).

The main advantage of the Theil index is that its value is calculated directly from the given time series. It is worth to stress that Theil index (Eq.(5)) transforms the time series into entropy time series (dependent on the time window size) than the distance between time series can be calculated by any distance measures (UD, MD or MMD). In the examples in Sec.4 the time evolution of the correlation between stock market prices of chosen companies is analysed. Since there is no reasons to expect linear correlation among entropy time series in the following analysis the Theil index based Manhattan Distance (ThMD) i.e. composition of the Theil index and MMD is used and the results compared with the standard UD analysis. (The interested reader can find the comparison of the results obtained by application of UD and MMD to the entropy transformed data in the case of macroeconomy data in [15].) The definition of ThMD is given by Eq. (6).

$$ThMD(A,B)_T = MMD(Th(A,T),Th(B,T))$$
(6)

It is important to stress that the Theil index require to define a time window, therefore the application of socalled moving time window analysis in the case of ThMD needs to define two different time windows – one to calculate the Theil index, the second to find the distance between time series in a given time window.

3. Distance measures properties

There is no reason to constrain the correlation analysis only to the case of the linear function. Therefore within this paper the general form is assumed (Eq.(7)). Let denote the time series as A and B and its elements as a_i and b_i respectively. The correlation between time series is understood as the existence of a reversible map with a one to one correspondence (an injection). Then the correlation between the time series A and B can be described as:

$$f: A \to B. \tag{7}$$

For the sake of correlation comparison the strength of correlation is introduced. If two functions are considered: $f_1: A \to B$ and $f_2: A \to C$, then the correlation is stronger if the same changes in origin time series (here A) results in bigger changes in correlated time series (here B or C).

The UD is based on the linear correlation coefficient and it gives negative results in the case of non-linear correlations. The above statement is illustrated by analytical and numerical examples.

Let X will be a random variable with finite variance and probability distribution function f(x) symmetrical with respect to the mean value i.e. $f(x) = f(-x), x \in$ $(-\infty, \infty)$. Then introducing a new random variable Y =|X| one get corr(X, Y) = 0, which is contradictory to the assumption.

Similar results are obtained if the correlation function Eq.(7) is given by a polynomial. Let consider three cases: the time series A is generated according to the formula: $a_i = i + w(0.5)$ then the time series are transformed by non-linear functions $f_1(A) = A^2 + w(0.5)$, $f_2(A) = A^3 + w(0.5)$, $f_3(A) = A^4 + w(0.5)$ and finally the UD calculated for the pairs: $g_1(n) = UD(A, f_1(A))$, $g_2(n) = UD(A, f_2(A))$, $g_3(n) = UD(A, f_3(A))$. The results are presented in Fig. 1. The main observation is that the distance between generated time series is far from zero, so it might be difficult to support the statement that the correlation exists. On the other hand if one consider the UD distance Eq.(2) as a measure of the correlation strength, than the result of the simulation is even more embarrassing, because the asymptotic results are as follows: $g_1 \rightarrow 0.125, g_2 \rightarrow 0.26, g_3 \rightarrow 0.2$, which means that in terms of UD the strongest correlation is observed in the f_1 case, which is the lowest degree polynomial considered here. Besides pointing the weakest correlation as the strongest UD confuses also the remaining functions since f_1 is the strongest, f_3 is weaker, and f_2 is the weakest. So it does not preserve the polynomial order.



Fig. 1. UD in the case of nonlinearly correlated time series: $g_1(n) = UD(A, B)$ where $a_i = i + w(0.5), b_i =$ $a_i^2 + w(0.5); g_2(n) = UD(A, B)$ where $a_i = i + w(0.5),$ $b_i = a_i^3 + w(0.5); g_3(n) = UD(A, B)$ where $a_i = i + i$ $w(0.5), b_i = a_i^4 + w(0.5).$

From the application point of view another important question is the problem of the noise resistance of the distance measure. Let us assume that a time series is influenced by a noise W with the mean value equal to zero

$$\langle W \rangle = 0. \tag{8}$$

The noise may influence the correlation either in A, B or on both time series. Then Eq.(7) takes one of the following forms:

$$f(A) = B + W, (9)$$

$$B = f^{-1}(A) + W (10)$$

or

$$f(A+W) = B+W \tag{11}$$

Eqs. (9)-(11) reflect the dependence problem, which is in general not trivial and should be treated with caution especially in the case of non-linear correlations, because the reverse function not always will exist or may have some strong constrains. Therefore the problem may be split into two cases: searching for linear or non-linear dependencies. In the following the noise influence on MD will be discussed in the case of correlation function with well defined reverse function f^{-1} then the analysis performed in the following section are valid in all three cases: Eqs.(9) - (11). Since the UD is defined to measure linear correlations the noise influence will be discussed in the case of linear correlation function f.

Assume that the white noise is present in both time series

$$A = A + W_A, \ B = B + W_B \tag{12}$$

By direct calculations it can be verified that the UD reads:

$$UD(A,B)_{(t,T)} = \frac{1}{\sqrt{2}} \times \tag{13}$$

$$\sqrt{1 - \frac{\langle AB \rangle - \langle A \rangle \langle B \rangle}{\sqrt{(\langle A^2 \rangle + \langle W_A^2 \rangle - \langle A \rangle^2)(\langle B^2 \rangle + \langle W_B^2 \rangle - \langle B \rangle^2)}}$$

The main feature of the Eq.(13) is that despite assumption of linearly correlated time series the noise W influences the calculated distance. The noise is present at Eq.(13) as the mean value of square $\langle W^2 \rangle$, which in general is not vanishing. The problem of noise influence on the UD is illustrated in Fig. 2. The following cases are considered: two time series with the same mean values, two time series with different mean values and time series linearly correlated $(b_i = 5 * a_i + w(0.5))$. The results are as follows: in the first two cases i.e. time series with constant mean value the distance between them is ≈ 0.7 , which suggests that there are no correlations. The result is somehow surprising because one could expect that the distance measure will find easily seen feature. Moreover UD in the third case of linearly correlated time series is also different from zero, which support the observation that the noise may hide existing correlations (see the conclusion of Eq.(13)).



= UD(A, B) where $\langle A \rangle = 0.5$, $\langle B \rangle = 0.5$; $f_2(n)$ UD(A, B) where $\langle A \rangle = 0.5$, $\langle B \rangle = 0.0$; $f_3(n)$ -UD(A, B) where $\langle A \rangle = 0.5, b_i = 5 * a_i + w(0.5).$

The noise influence on the results of MD is considered in the case of time series A, B which satisfy the condition $\forall i \ a_i > b_i$, then $MD(A, B) = \sum_{i=1}^n |a_i - f^{-1}(a_i) - w_i| = \sum_{i=1}^n (a_i - f^{-1}(a_i) - w_i) = \sum_{i=1}^n (a_i - f(a_i))$. This shows that the Manhattan Distance in some situations can be robust to a noise. The same results is obtained in the case: $\forall i \ a_i < b_i$. From the practical point of view economy data due to the differences in size of entities very often satisfy this condition, which make MD very useful.

4. Network analysis of distance matrices

Most of the correlations analyses are performed to discover structure of the network between companies or countries [22–27]. The key point is that the distance matrix obtained by application of any of distance measures is difficult to analyse directly. Therefore various network methods are applied. In the following the most popular Minimum Spanning Tree (MST) and its modification Bidirectional Minimum Length Path (BMLP) and Unidirectional Minimal Length Path (UMLP) will be used. The algorithms generating MST can be found in various textbooks e.g. [28]. The best known are Prim's and Kruskal's algorithms [29, 30]. Definitions of BMLP and UMLP can be found in [7, 8, 31, 32]. For the convenience of the reader the key elements of the algorithms will be recalled. In the case of Kruskal algorithm of MST one has to find a shortest edge and add repeatedly the next shortest edge that does not produce a cycle. The algorithm stops when there is a path between any two nodes. BMLP can be considered as a simplification of MST algorithm. First the pair of the closest neighbours is searched for and becomes the seed with the first two ends of the chain. Then the two nodes closest to each end of the chain are searched for, and the one with the shortest distance to one of the ends is attached and becomes the new end on that side of the chain. The procedure is repeated. The chain grows in two directions. The BMLP graph does not contain loops, the node degree does not exceed two, i.e. there are no nodes with more than two neighbours. In the case of UMLP the algorithms begins with the choice of the seed of the network. Then the closest item to the end of the chain is searched for and attached becoming the end of the network. Nodes are attached only to the end of the network. Therefore it grows only in one direction. As in the case of BMLP there are no loops. The main advantage of UMLP is the possibility to choose the seed of the network since in various analyses there is such a "natural" reference point e.g. stocks market index, or a leader in a given group.

In the following examples two aspects will be illustrated: the network analysis of distance matrices and the comparison between the ultrametric and entropy based distances. The network analysis of distance matrices obtained for UD and ThMD will be performed for evolving network in the case of MST, BMLP and UMLP. The following two groups of companies will be analysed:

• WIG20: PEKAO, PKO BP, KGHM, PKN ORLEN, TPSA, BZ WBK, ASSECO POLAND, CEZ, GETIN HOLDING, GTC, TVN, PBG, POLIMEXMS, BRE, LOTOS, CYFROWY POL-SAT, BIOTON. The considered time period extends from 05.01.2009 to 30.04.2010, thereby including a

05.01.2009 to 30.04.2010, thereby including a total of 334 quotes. The closing prices are considered in the analysis.

• S&P 500: ABB Ltd.(ABB), Apple Inc. (AAPL), Boeing Co. (BA), the Coca-Cola Company (KO), Emerson Electric Co. (EMR), General Electric Co. (GE), Hewlett-Packard Company (HPQ), Hitachi Ltd. (HIT), IBM (IBM), Intel Corporation (INTC), Johnson & Johnson (JNJ), Lockheed Martin Corporation (LMT), Microsoft Co. (MSFT), Northrop Grumman Corporation (NOC), Novartis AG (NVS), Colgate-Palmolive Co. (CL), Pepsico Inc. (PEP), Procter & Gamble Co. (PG), Tower Semiconductor LTD. (TSEM), Wisconsin Energy Corporation Co. (WEC). In the analysis the closing prices are used. The considered period: 02.01.2009 – 30.04.2010, which gives 334 price quotes per stock in total. Data were obtained from the web page: finance.yahoo.com.

The main advantage of UD is ability to distinguish clusters of companies formed due to the industry sectors. This ability is illustrated in Fig. 3. In the generated MST the following industry clusters can be distinguished:

- food industry: PG, PEP, KO, JNJ
- computer and electronic sector: AAPL, HPQ and INTC, TSEM, MSFT
- aircraft industry: BA, NOC, LMT.

An additional outcome is the analysis of the network degree i.e. the number of connections to the given node, which can point out potential leaders here it is EMR, which is a diversified global technology company.



Fig. 3. MST network generated from the distance matrix of UD in the case of chosen S&P 500 companies..

Besides the static (at a given moment of time) analysis of the correlation very often the evolution of the system is investigated. In such a case the analysis requires construction of huge number of networks. Although it might be possible to analyse them in the same manner as in the later case an alternative approach is presented. The attention of the reader is directed to another aspect of the problem – the analysis of correlations of the given group. The following parameters are analysed: the mean distance between companies on the network and the results compared with respect to the type of the distance applied and the network constructed. The following time windows are chosen: T = 100 days and $T_1 = 50$ days, $T_2 = 50$ days in the case of UD and ThMD respectively. The results are presented in Fig. 4 and Fig. 5 in the



Fig. 4. The mean distance between the companies of the chosen subset of WIG20 index in the case of MST, BMLP and UMLP networks for the time widow T = 100 days and $T_1 = 50$ days, $T_2 = 50$ days in the case of UD and ThMD respectively.



Fig. 5. The mean distance between the companies of the chosen subset of S&P500 index in the case of MST, BMLP and UMLP networks for the time widow T = 100 days and $T_1 = 50$ days, $T_2 = 50$ days in the case of UD and ThMD respectively.

case of subset companies quoted in WIG20 and S&P 500 respectively. The most striking observation is that the time evolution of the mean distance between companies depends on the distance measure applied. The second feature is that evolution of the mean distance between the nodes in the case of the ThMD is "richer" than in UD and allows pointing out several periods increases of correlations as well as decreases. In fact ThMD measures difference of complexity of the time series in the information theory sense, therefore one can speculate that ThMD reflects changes in general knowledge of the stock markets players. Another important remark refers to the types of the network applied. In all considered cases the less "noisy" results are obtained for MST, however the outcome of other networks i.e. BMLP and UMLP do not differ significantly from the MST results. For example in Fig. 4 the mean distance of MST and BMLP in the case of UD almost coincide. The same situation is observed for S&P500 BMPL and UMLP networks in the case of UD. In the case of ThMD such a close overlap is not observed. However, the mean distance functions have the extreme points at similar moment of time. The application of different network types do not alter the main results of the analysis i.e. position of the extreme points as well as the periods of increase or decrease of the mean distance are preserved.

Summarising the above observations it can be stated that application of entropy based distance measure gives significantly different information and its application is worth consideration if the main goal of the analysis is different than portfolio optimisation. The second important result is that networks which require significantly less computational power gives comparable results. Therefore the BMLP and UMLP networks might be a good choice in the system of real time computations where time series of significant length have to be preceded.

5. Conclusions

The main objective of the study was the presentation of the time series correlation distance measures and the distance matrix analysis. Four different distance measures were presented, which can be divided into two groups: the correlation measures (UD) and Manhattan distance (MD) with its modifications MMD, ThMD. The main advantages of UD are:

- It follows from the theory of optimal portfolio.
- Properly classify entities in the context of portfolio optimization.
- Verifies the linear correlations.

Among the disadvantages of UD the most important are:

- It is sensitive to the presence of noise.
- It fails to compare correlations of different types i.e. different than linear.

The proposed alternative correlation distance measure was MD. The main advantages of MD are:

- In the case of time series of significantly different values it is not sensitive to the noise with the mean value equal to zero.
- It allows comparing different types of correlations.
- It makes possible to introduce correlation classes.

The main disadvantage of MD is:

• It measure only mutual correlation and do not allow to decide which company is the leading one.

Besides directly measured distances the entropy based distances were discussed. Although in the literature the best known entropy measure is the Shannon entropy this formulation was not used due to the difficulties in application. As an alternative entropy measure the Theil index was chosen and its features illustrated on the real data examples. Additionally the network analysis of the distance matrices was discussed especially the application of different network type in the case of real data systems. References

- S.B. Goldenberg, L.J. Shapiro, J Climate 9, 1169 (1996).
- [2] M. Pascual, X. Rodo, S.P. Ellner, R. Colwell, M.J. Bouma, *Science* 289, 1766 (2000).
- [3] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A. Nunes Amaral, H.E. Stanley, *Phys. Rev. Lett.* 83, 1471 (1999).
- [4] T. Aste, T. Di Matteo, M. Tumminello, R.N. Mantegna, in: Noise and Fluctuations in Econophysics and Finance, Ed. D. Abbott, J.-P. Bouchaud, X. Gabaix, J.L. McCauley, Vol. 5848 SPIE, Bellingham, WA 2005, p. 100.
- [5] A. Razdan, *Phys A* **333**, 335 (2004).
- [6] V. Tola, F. Lillo, M. Gallegati, R.N. Mantegna, J. Econ. Dyn. Control 32, 235 (2005).
- [7] J. Miśkiewicz, M. Ausloos, Acta Phys. Pol. B 36, 2477 (2005).
- [8] M. Ausloos, J. Miśkiewicz, in *Practical Fruits of Econophysics*, Ed. H. Takayasu Springer, Tokyo 2006.
- [9] R.N. Mantegna, H.E. Stanley, An introduction to econophysics: correlations and complexity in finance, Cambridge University Press 2000.
- [10] H. Markowitz, H.M. Markowitz, Portfolio selection: efficient diversification of investments, Wiley-Blackwell 1991.
- [11] G. Bonanno, G. Caldarelli, F. Lillo, R.N. Mantegna, *Phys. Rev. E* 68, 046130 (2003).
- [12] G. Bonanno, N. Vandewalle, R.N. Mantegna, *Phys. Rev. E* 62, R7615 (2000).
- [13] J.P. Onnela, A. Chakraborti, K. Kaski, J. Kertesz, *Phys. A* **324**, 247 (2003).
- [14] B.M. Roehner, *EPJ B* **13**, 175 (2000).

- [15] J. Miśkiewicz, M. Ausloos, Phys. A 389, 797 (2010).
- [16] S. Willard, *General Topology*, Dover Publications 2004.
- [17] M. Gligor, M. Ausloos, J. Econ. Integration 23, 297 (2008).
- [18] F. Chybalski, E. Marcinkiewicz, I. Staniec, Polish J. of Environ. Stud. 18, 68 (2009).
- [19] J. Miśkiewicz, *Phys. A* **387**, 6595 (2008).
- [20] T. Downarowicz, *Entropy in Dynamical Systems*, Cambridge University Press 2011.
- [21] H. Theil, *Econometrica* **33**, 67 (1965).
- [22] G. Bonanno, F. Lillo, R.N. Mantegna, *Phys. A* 299, 16 (2001).
- [23] G. Bonanno, G. Caldarell, F. Lillo, S. Micciche, N. Vandewalle, R.N. Mantegna, *Eur. Phys. J. B* 38, 363 (2004).
- [24] F. Lillo, R.N. Mantegna, *Phys. A* **338**, 125 (2004).
- [25] T. Di Matteo, T. Aste, R.N. Mantegna, Phys. A 339, 181 (2004).
- [26] T. Di Matteo, T. Aste, in Abstract Book of the XXIII IUPAP International Conference on Statistical Physics, Ed. L. Pietronero, V. Loreto, S. Zapperi 2007.
- [27] F.O. Redelico, A.N. Proto, P. Clippe, M. Ausloos, J. Phys. Conf. Ser. 221, 012001 (2010).
- [28] J.A. Bondy, U.S.R. Murty, Graph theory, Springer 2008.
- [29] J.B. Kruskal, in: Proceedings of the American Mathematical Society, 7, 48 (1957).
- [30] R.C. Prim, AT&T Tech J 36, 1389 (1957).
- [31] J. Miśkiewicz, M. Ausloos, Int. J. Mod. Phys. C 17, 317 (2006).
- [32] M. Ausloos, J. Miśkiewicz, Int. J. Bifurcation and Chaos 20, 381 (2010).