Proceedings of the 8th Polish Symposium of Physics in Economy and Social Sciences FENS, Rzeszów, November 4–6, 2015

Multiway Similarity Approach Based on Divergence Functions and Smoothness Measure

R. Szupiluk* and T. Sobon
Warsaw School of Economics, al. Niepodległości 162, 02-554 Warsaw, Poland

In this paper we present a novel similarity measure method for financial data. In our approach, we propose the assessment of the similarity in a coherent hierarchical and multi-faceted way, following the general scheme where various detailed basic measures may be used like the Fermi–Dirac divergence, Bose–Einstein divergence, or our new smoothness measure. The presented method is tested on benchmark and real stock markets data.

DOI: 10.12693/APhysPolA.129.927
PACS/topics: 05.45.Tp, 05.40.Ca, 07.05.Kf, 07.05.Mh

1. Introduction

The assessment of similarity is one of the fundamental issues occurring in the analyses, models and theories of financial markets. In particular, it is a significant element of the investment process where fundamental economic and financial analysis is not the key factor of price shaping. This is also the case in trading systems especially for high frequency trading (HFT), where fundamental factors are generally stable during the investing actions. The popularity of such investing approaches does not mean that the similarity problem is clear or solved. On the contrary, we can say that generally the problems with market similarities, patterns and associated with them transaction rules lead to hot discussions on how to measure, detect or recognize similar stock formations and their connections to assumed patterns [1]. The assessment of similarity is one of those skills of a human mind that we use quite frequently and efficiently, but it is not easy to describe, explain, or define it precisely. For example, while in many cases it is relatively easy to see a family resemblance, a formal justification for such similarities is usually ambiguous and vague. We have an analogical situation in the case of financial time series, where human perception can be often far from formal approaches. Our aim is to find a method that will give results consistent with the human perception.

From the calculation perspective, we understand objects lying close to each other as similar but to some extent tight-knit/cohesive or interdependent. With regard to the similarity of time series we can speak about the direct similarity (distance) between the data as well as the distance between certain characteristics of the signals. Variation, smoothness and sparsity are most common among these characteristics. Although the concept of similarity as well as relationships, variability, smoothness can be defined, measured and interpreted differently, at the same time in many cases and views they also approach, penetrate and intertwine. For example, in case of the multi-dimensional variability calculated as the variance, it depends also on the value of covariance which is usually interpreted as a measure of dependence.

In practical applications we can distinguish two main approaches for the similarity measure. The first one is based on the basic or elemental measures like covariance, norms or divergences. The second approach uses complex systems of pattern exploration such as neural networks, decisions trees, or support vector machines (SVMs), which in many cases is a very effective solution. However, systems of pattern exploration are based on basic measures and their construction requires a certain learning process based on prototypes. Therefore, their effectiveness can be treated as a derivative of the proper basic functions evaluation and good prototypes choice. It should also be noted that complex systems act often like black boxes that may be successfully applied for individual solutions but their theoretical value is limited.

In our concept we propose an innovative method which can be treated as a compromise between systems of pattern exploration and basis/elemental measures. We assess the similarity in a coherent hierarchical and multi-faceted way, following the defined general scheme where various detailed basic measures may be used, such as second order statistics (SOS) and higher order statistics (HOS) covariances, norms or divergence functions. On the level of such basis measures we propose our innovative RS divergence concept, which can be connected with well known Fermi–Dirac and Bose–Einstein divergences. This approach is specially addressed for direct price analysis, where signals are nonnegative. Of course, the method can be applied for any data after simple preprocessing to non-negativity.

2. Similarity and second order statistics — limitation and inspiration

Standard quantitative similarity measures are based on terms which are tangible and can be easy for mathematical interpretation, such as correlation or metric dis-

*corresponding author; e-mail: rszupi@sgh.waw.pl

tance. From such calculation perspective the central role is played by the SOS. The knowledge of the second-order statistics completely defines the parameters of the normal distribution which provides full statistical information about the phenomenon. In conjunction with the role of the normal distribution in linear models with stationary signals we get an elegant mathematical instrument that dominates in data analysis in recent years. However, these successes are largely related to the characteristics of useful technical and natural signals, which in terms of a SOS paradigm significantly differ from the characteristics of noise signals [2, 3]. In the case of economics and finance, the situation is quite different. Typical financial signals such as stock prices are nonstationary and their rate of return has often characteristics close to white noise [4, 5]. In both cases the use of correlation methods to identify, distinguish or compare both of the instruments as well as random noise separation are significantly impeded. Additionally, even very similar signals in human sense can be independent (dissimilar) in a SOS sense. But similarity, assessing by covariance and variance, can be motivation for our proposition. Let us note that if we calculate similarity as a covariance value, it means that in epistemological terms similarity is connected with dependence, similarity:=dependence. But in the multivariate case the total variance is expressed as a sum of variance and covariance.

$$\overbrace{\operatorname{var}(x_1+x_2+...+x_m)}^{\text{variability}(A)} = \overbrace{\sum_{i=1}^{m}\operatorname{var}(x_i)}^{\text{variability}(B)} + \overbrace{2\sum_{i=1,j=1,i\neq j}^{m}\operatorname{cov}(x_i,x_j)}^{\text{dependence}}. \quad (1)$$

This means that the concepts of variability and dependence are related and for (1) it can be written as $\text{variability}_A - \text{variability}_B \overset{*}{\approx}{}^{\dagger}$ dependence. The final conclusion leads to a relation where similarity $\overset{*}{\approx}$ variability. Such epistemological terms properties are used in our approach. But to avoid direct connections variance with its assumptions, methodology and standard interpretations, smoothness signal characteristics are explored. Of course, smoothness and variability by variable in many situations explore similar characteristics of signals. In practice its meaning and interpretation depends on the adopted definition and applied quantitative measures.

3. Assessment of the similarity the first step — elementary functions and variability

We start our similarity assessment approach by constructing smoothness characteristics that could be the basis function for the general scheme. Our aim is to create a characteristic which is directly associated with the nature of financial processes. For this purpose, we explore the ideas of the divergence functions [6–8]. Divergence $D(y\|z)$ is a function, defined on non-negative variables z and y, that satisfies $D(y\|z) \geq 0$ and $D(y\|z) = 0$ only if $y = z$. Divergence does not need to satisfy the triangle inequality $D(y\|z) \leq D(y\|x) + D(x\|z)$ and the condition of symmetry does not have to be met for it, which means that divergence is usually asymmetric $D(y\|x) \neq D(x\|y)$. Divergences are interpreted as a measure of differentiation, quasidistances or differences. Due to its nonnegativity, the following assumptions can be addressed for directly exploring stock prices. For measuring the smoothness between the sequence $y(t)$ and $y(t-k)$ we define the autodivergence function as $D^k(y) = D(y(t)\|y(t-k))$. Such formulated autodivergence function has the same properties like standard divergence function but it is addressed for single signal. In further considerations we focus on the Fermi–Dirac divergence and Bose–Einstein divergence described in the modified form as:

a) Fermi–Dirac autodivergence

$$D^k_{\text{FD}}(y) = \sum_t \Big[y(t)\ln\frac{y(t)}{y(t-k)} + (1-y(t))\ln\frac{1-y(t)}{1-y(t-k)}\Big], \quad (2)$$

for $y_i, z_i \in [0,1]$; b) Bose–Einstein autodivergence

$$D^{\alpha,k}_{\text{BE}}(y) = \sum_t y_t \ln\frac{(1+\alpha)y(t)}{y(t)+\alpha y(t-k)} + \alpha y(t-k)\ln\frac{(1+\alpha)y(t-k)}{y(t)+\alpha y(t-k)}. \quad (3)$$

Now, we introduce a new divergence which explores fundamental financial characteristics such as normal and logarithmic rate of returns. We define the measure calculated as a sum of absolute values of normal and scaled logarithmic returns as

$$D^k_{\text{SR}}(y) = \sum_t \Big|y(t) - y(t-k) + 0.5(y(t) + y(t-k))\ln\frac{y(t)}{y(t-k)}\Big|. \quad (4)$$

The scaling operation with $0.5[y(t) + y(t-1)]$ is performed for consistency of different types of returns. The formulae (4) can be interpreted as an effect of the point-to-point shifts in the phase space. We can see an interesting connection of D_{SR} divergence with D_{FD} and D_{BE} divergences. For $\alpha \in (0,1)$ in D^{α}_{BE} we can make approximation $\ln y \approx y - 1$ to obtain $D^{\alpha}_{\text{BE}}(y) = \sum_i \alpha(y(t) - y(t-k))^2 + R_B = \sum_i f_{\text{BE}}(y(t), y(t-k)) + R_B$ where R_B means residuals associated with approximation precision. After an analogical approximation for D_{FD} we obtain $D_{\text{FD}}(y) = \sum_t (2y(t)^2 - 2y(t)y(t-1) + y(t-1) - y(t)) + R_F = \sum_t f_{\text{FD}}(y(t), y(t-k)) + R_F$. Leaving approximation residuals for $\alpha \approx 1$ we have $D_{\text{SR}} \approx \sum_t |f_{\text{BE}} - f_{\text{FD}}|$ which allows us to intemperate of (4) in a deep and elegant statistical way [9]. Dependences presented above describe some basis characteristics of signals, in our case interpreted in terms of smoothness. Currently, we pro-

† $\overset{*}{\approx}$ means described or defined.

pose a concept of a multi-way comparison of signals using such characteristics. The starting point for our considerations is the case of Gaussian signals, which are similar not only from an analytical point of view but also taking into account the human perception. The crucial property of such signals is that their sum is also a Gaussian signal. Consequently, the average smoothness of a single Gaussian signal is equal to the smoothness of the sum of those signals. Moreover, those dependences do not depend on neither the delay parameter k nor the sequence of the arguments. Taking into account the above properties we may now define the measure of the similarity between signals as:

$$\Phi(y_1, y_2) = a\frac{D(y_1 + y_2)}{D(y_1) + D(y_2)}, \tag{5}$$

where $D(y)$ is chosen as smoothness measure and a is a scaling factor. Such similarity measure is associated with some reference similarity pattern given by Gaussian signals. It can be interpreted in terms of the signal non-gaussianity. Of course, just like other similarity criteria, it is in some sense an arbitrary assumed convention "what similarity mean". Formula (5) allows us to determine the mutual relationship (similarity) between the two signals based on individual characteristics of smoothness. In the multivariate case based on (5) we can create a similarity matrix form

$$\Phi = \Phi(y_n, y_m)_{mn}. \tag{6}$$

Starting from the basic idea of assessing the similarity-based formula (6), we can define further relations allowing for the analysis of the overall similarity of many variables, the similarities between variable and groups of variables or similarity between different groups. We present several examples for different cases.

a) The mutual similarity of three single signals in our method is measured as

$$\Phi_{[1,2,3]}(y_1, y_2, y_3) = a\frac{D(y_1 + y_2 + y_3)}{D(y_1) + D(y_2) + D(y_3)}. \tag{7}$$

b) Analysis of the similarity of the group of variables

$$\Phi_{[1,2,3,4,5]:[1,2],[3,4,5]}(y_1, y_2, ..., y_5) =$$

$$a\frac{D(y_1 + y_2 + ... + y_5)}{D(y_1 + y_2) + D(y_3 + y_4 + y_5)}. \tag{8}$$

The further extension of system similarity analysis is to introduce general dividing signals relations, in other word analysis of the impact of the group of variables division (or all variables) into sub-groups. For example

$$\Phi_{[1,2,3,4,5]:[1],[2,3,4,5]}(y_1, y_2, ..., y_5) =$$

$$a\frac{D(y_1) + D(y_2 + y_3 + y_4 + y_5)}{D(y_1 + y_2) + D(y_3 + y_4 + y_5)}. \tag{9}$$

Selecting the proper characteristics should be related to the objective of the analysis. If we are interested in the analysis of the interrelationships the views from point (7) seem to be appropriate. If we analyze the variation of the variable part (or groups) relative to the other variables, items (8) and (9) are the right ones.

4. Practical experiment

In this paragraph, we present the aforementioned above conception idea in practical tests. We start with the main question which is aimed at looking for a good and "intuitive" quantitative measure of similarity.

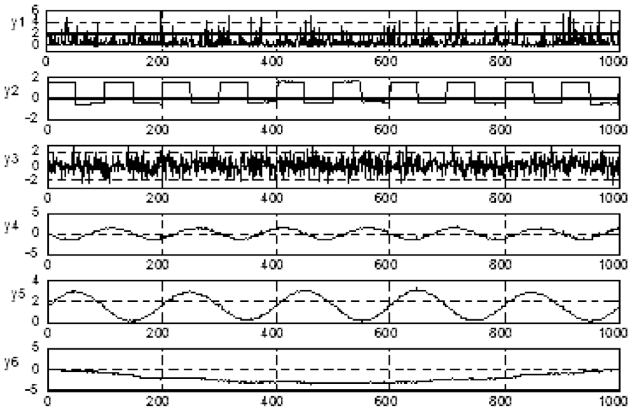

Fig. 1. Signals tested for similarity problem.

Figure 1 presents six signals which are independent (and thus decorrelated) with individual unity variances, and therefore we have no information from SOS analysis. Moreover, it can be seen that the p-norm distance calculated as $\|x\|_p = (\sum_{i=1}^{m} |x_i|^p)^{1/p}$, given in Table I, does not correspond to the visual evaluation.

The signal comparing of the signals via the smoothness measure by divergences gives preliminary information according to the difference between those characteristics. The smoothest signals are y_5 and y_6, whereas the least smooth signals are noise signals: y_1 and y_3. The differences between various autodivergence measures are presented in Table II.

TABLE I

The distances between the signals measured with p-norm, for $p = 2$.

$\times 10^{-4}$	y_1	y_2	y_3	y_4	y_5	y_6
y_1	0	483	411	452	452	452
y_2	483	0	292	348	348	347
y_3	411	292	0	239	239	238
y_4	452	348	239	0	304	301
y_5	452	348	239	304	0	304
y_6	452	347	238	301	304	0

TABLE II

The signal smoothness.

$\times 10^{-4}$	y_1	y_2	y_3	y_4	y_5	y_6
$D_{\mathrm{BE}}^{0.5;1}$	306	124	422	35	4	16
D_{FD}^{1}	9637	553	2987	122	19	297
D_{SR}^{1}	590	236	1832	275	113	181

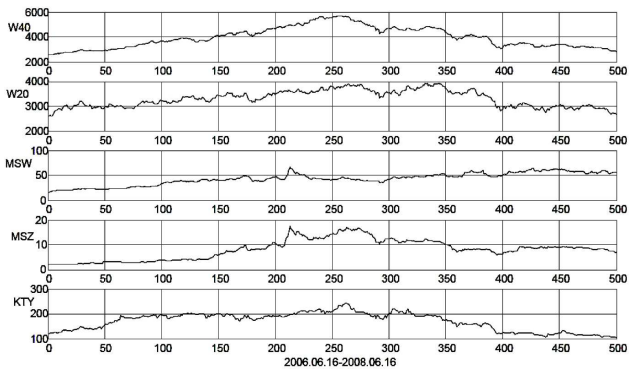

Fig. 2. Real financial time series.

On the other hand, the similarity measure represented by matrix Φ, indicates the biggest similarity between signals: y_1 and y_3, y_4 and y_5, y_2 and y_4. It is presented in Table III. Signal y_6 is different from the others. It is crucial that noise signals are significantly different from deterministic signals.

In this case, our measure of similarity of benchmarked signals is compatible with our intuition. This basic test is an important step before assessing the similarity of much more complicated financial time series.

Another experiment was done using real financial time series of indexes mWIG40, WIG20 and stocks from the Warsaw Stock Exchange: Mostostal Warszawa (MSW), Mostostal Zabrze (MSZ) and Kety (KTY). The data is presented in Fig. 2. We are interested in verifying whether or not our method gives the same results as the "human insight".

TABLE III

The signal similarity measured via matrix Φ for D_{SR}.

	y_1	y_2	y_3	y_4	y_5	y_6
y_1	1	1.4495	0.9526	1.4726	1.437	1,2561
y_2	1.4495	1	1.43	0.8958	0.7972	0.8004
y_3	0.9526	1.43	1	1.3267	1.4191	1.3436
y_4	1.4726	0.8958	1.3267	1	1.108	0.4794
y_5	1.437	0.7972	1.4191	1.108	1	0.6553
y_6	1.2561	0.8004	1.3436	0.4794	0.6553	1

TABLE IV

The real financial signal similarity measured via matrix Φ for D_{SR}.

	y_1	y_2	y_3	y_4	y_5
y_1	1.0002	0.8113	0.8505	0.8755	0.7079
y_2	0.8113	1.0003	1.1811	1.2333	1.0287
y_3	0.8505	1.1811	1.0000	0.9047	1.2717
y_4	0.8755	1.2333	0.9047	1.0003	1.2407
y_5	0.7079	1.0287	1.2717	1.2407	1.0005

TABLE V

The real financial signal smoothness measured in particular measures.

$\times 10^{-4}$	y_1	y_2	y_3	y_4	y_5
$D_{\mathrm{BE}}^{0.5;1}$	3.3	13.3	7.1	5.7	12.4
D_{FD}^{1}	24.7	116.3	21.4	34.0	108.0
D_{SR}^{1}	572.0	1200.1	775.8	733.0	1067.4

Table IV presents matrix Φ for real financial time series. The most similar are y_3 (MSW) and y_4 (MSZ). It is consistent with the human insight as well as with the market characteristics as since both stocks come from the same construction branch — construction. Generally, all above financial time series are similar, which might be caused by the fact that the data comes from the same time period. That time was characterized by strong growth of the whole market and after that most stocks decreased. From a quantitative point of view, it means a strong correlation of most financial instruments.

The outcomes coming resulting from matrix Φ are consistent with the direct smoothness analysis presented in Table V.

We can observe the similar smoothness of stocks from the construction branch. Such observations can give us direct information for the selection of instruments in the APT theory or can become a basis for the construction of the AT trading system.

5. Conclusions

Our method allows to assess the similarity between the individual signals as well as their groups. Such flexible approach gives results, which are close to human perception which we show on benchmark signals. We can see that our method can identify similar signals in case, when standard correlation methods can fail. After choosing of Φ function type, the similarity measure has close analytical form and allows the direct calculation without any learning processes (optimization) on specific prototypes, which is typical for methods of artificial intelligence. The Φ function create a general assessment scheme, where different basis function can be used. We introduce a new divergence/autodivergence function motivated by typical financial difference calculations. Such function can be connected with the Fermi–Dirac and Bose–Einstein divergences, which allows for deep theoretical interpretations. Practical experiments with real market data confirm validity of our approach. In our opinion this method can be straightly used in APT theory for instruments selection, which is our task for further research as well as testing the full system with others basis measures.

References

[1] J. Krutsinger, *Trading Systems: Secrets of the Masters*, McGraw-Hill, New York 1997.

[2] J.L. Rodgers, W.A. Nicewander, *Am. Statistic.* **42**, 59 (1988).

[3] C.W. Therrien, *Discrete Random Signals and Statistical Signal Processing*, Prentice Hall, New Jersey 1992.

[4] A.N. Shiryaev, *Essentials of Stochastic Finance: Facts, Models, Theory*, World Sci., Singapore 1999.

[5] R. Szupiluk, P. Wojewnik, T. Zabkowski, *Noise Detection for Ensemble Methods, Lecture Notes in Artificial Intelligence*, Vol. 6113, Springer, Heidelberg 2010.

[6] S. Amari, *Diferential-Geometrical Methods in Statistics*, Springer Verlag, New York 1985.

[7] A. Cichocki, R. Zdunek, A.-H. Phan, S. Amari, *Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis*, Wiley, Tokyo 2009.

[8] I. Csiszar, in: *Prague Conf. on Information Theory*, Vol. A, Academia, Prague 1974, p. 73.

[9] L. Knockaert, *IEEE Trans. Sign. Process.* **41**, 3171 (1993).