METODY ILOŚCIOWE W BADANIACH EKONOMICZNYCH
Tom XIII/3, 2012, str. 213 – 221

DYWERGENCJE BOSEGO-EINSTEINA W ANALIZIE PODOBIEŃSTW FINANSOWYCH SZEREGÓW CZASOWYCH[1]

Ryszard Szupiluk
Katedra Informatyki Gospodarczej
Szkoła Główna Handlowa w Warszawie
e-mail: rszupi@sgh.waw.pl

Streszczenie: Ocena wzajemnego podobieństwa finansowych szeregów czasowych jest jednym z problemów, w którym kwestia właściwego doboru metody analitycznej zaznacza się bardzo wyraźnie. Z reguły problem ten sprowadzany jest do analizy korelacji - co nie zawsze prowadzi do właściwych rezultatów. Często są to oceny wręcz sprzeczne ze wizualną obserwacją lub wiedzą ekspercką. Powodów takiego stanu rzeczy można upatrywać zarówno we właściwościach samej miary korelacyjnej i jej adekwatności do analizowanych danych, jak również w aspekcie metodologicznym przeprowadzanego badania. W niniejszym artykule zaproponujemy alternatywne rozwiązanie oparte na miarach dywergencji, w szczególności dywergencji Bosego-Einsteina. Przeprowadzone eksperymenty na poglądowych danych symulowanych potwierdzają użyteczność zaproponowanych rozwiązań

Słowa kluczowe: podobieństwo szeregów czasowych, miary dywergencji, dywergencja Bosego-Einsteina

WSTĘP

Podobieństwo między zmiennymi, wektorami czy funkcjami może być różnie definiowane. W przypadku analizy i modelowania danych empirycznych, w rzeczywistych problemach ekonomicznych, często oczekuje się by matematyczne ilościowe oceny podobieństwa odpowiadały potocznej intuicji z tym pojęciem wiązanej. W takim rozumieniu problematyka oceny podobieństwa

[1] Projekt został sfinansowany ze środków Narodowego Centrum Nauki przyznanych na podstawie decyzji numer DEC-2011/03/B/HS4/05092

instrumentów finansowych, reprezentowanych przez odpowiadające im szeregi czasowe, zasadniczo ujmowana jest w dwojaki sposób.

Najpopularniejszym, najszerzej zbadanym i opisanym podejściem są metody korelacyjne. Podobieństwo jest tu interpretowane w kategoriach statystycznych zależności drugiego rzędu. Podejście to ma długą tradycję, doprowadziło także do rozwiązania wielu fundamentalnych zagadnień, lub stanowiło podstawę tychże rozwiązań. Wiąże się to, z ogólnym znaczeniem statystyk drugiego rządu i rozkładu normalnego w analizie danych. Rozkład normalny, którego znaczenie statystyczne ugruntowane jest w centralnym twierdzeniu statystycznym, jest w pełni zdefiniowany przez statystyki drugiego rzędu. Stosunkowo prosta estymacja statystyk drugiego rzędu, w przypadku rozkładu gaussowskiego, pozwalana na uzyskanie pełnej informacji statystycznej o badanym zjawisku. Dodatkowo, na bazie rozkładu gaussowskiego, oraz dominującego paradygmatu estymacji, w postaci metody największej wiarygodności, można otrzymać „wygodne", liniowe modele statystyczne [10].

Istnieje jednak szereg zagadnień, gdzie wspomniany aparat korelacyjny nie jest tak adekwatny. Takie uwarunkowania jak: zależności nieliniowe, dane o klastrowej strukturze, niestacjonarność zmiennych, w przypadku mechanicznego zastosowania aparatu korelacyjnego, mogą prowadzić do niewłaściwych wniosków [2,9]. Transformacje zmiennych z kolei, mogą doprowadzić do zniekształcenia pierwotnych związków między zmiennymi. W efekcie, badanie zależności korelacyjnych, wymaga w dużej mierze indywidualnego nadzoru, co utrudnia wykorzystanie w automatycznych systemach rozpoznawania wzorców [7,8]. Także na gruncie samych metod statystycznych znajdziemy zagadnienia, kiedy wariancja jest z założenia niejednoznaczna. Jednym z takich przykładów są zastosowania analizy składowych niezależnych. Jest to wielowymiarowa metoda z założenia adresowana do zmiennych niegaussowskich (poza jedną dopuszczalną w zbiorze) oraz eksplorująca statystyki wyższych rzędów. Przy czym uzyskiwane komponenty w tej metodzie cechują się, niejednoznacznością względem wariancji. Oznacza to, że może być ona dowolnie skalowalna, co powoduje że wiele algorytmów realizujących ICA z założenia standaryzuje wariancje do jedności [3]. Przy czym, „wizualna" charakterystyka zmienności tych komponentów może być wysoce zróżnicowana.

Alternatywne podejście w ocenie podobieństwa, polega na postawieniu zagadnienia jako problemu segmentacji (grupowania). Prowadzi to szerokiego spektrum różnych technik w których podobieństwo oceniane jest z reguły jako odległość euklidesowa między badanymi obiektami, lub w ogólnym przypadku jako odległość mierzona określoną *p*-normą.

$$D_p\|\mathbf{x}-\mathbf{y}\|_p = \left(\sum_{i=1}^{L}|x_i - y_i|^p\right)^{1/p} \tag{1}$$

Jest to także popularne podejście, którego jednak głównym mankamentem jest, ponownie niejednoznaczność otrzymywanych wyników. W przypadku dwóch zmiennych (szeregów czasowych, sygnałów) z samej informacji o odległości między nimi, trudno jest ocenić stopień ich podobieństwa. Pewnym rozwiązaniem może być tu przyjęcie określonej zmiennej referencyjnej, względem której obliczane są odległości, jednak całokształt tak uzyskanej informacji o podobieństwie jest dalece niejednoznaczny i względny.

Powyższe ograniczenia, były i są motywacją do poszukiwania nowych miar podobieństwa. W ostatnim czasie szczególnie intensywnie badane są miary dywergencji. Rozwój metod opartych na dywergencjach, wiąże się przede wszystkim ze spektakularnymi sukcesami ich zastosowań, w zagadnieniu nieujemnej faktoryzacji macierzy, dzięki której, można z kolei dokonać niezwykłych operacji na obrazach i wzorcach, np. wyodrębnić pojedynczy jego fragment. Otworzyło to nowe możliwości budowy systemów rozpoznawania wzorców, m.in. dla komunikacji ze sparaliżowanymi, za pomocą analizy obrazów fal mózgowych. Jedną z cech charakterystycznych dywergencji, jest ich ogólny brak symetrii, co może zostać wykorzystane do oceny podobieństwa sygnałów (szeregów czasowych). W niniejszym artykule, zostanie zaproponowana metod oceny podobieństwa szeregów czasowych (sygnałów), oparta na dywergencji Bosego-Einsteina. Prezentacja tej metody zostanie dokona w kontekście filtracji szeregów czasowych przy wykorzystaniu analizy składowych niezależnych. Pozwoli to na wykazanie naturalnych ograniczeń związanych z analizą korelacyjną. Przeprowadzone eksperymenty praktyczne, zarówno na poglądowych danych symulowanych, jak i na rzeczywistych instrumentach finansowych potwierdzają użyteczność zaproponowanych rozwiązań.

DYWERGENCJE BOSEGO-EINSTEINA I PODOBIEŃSTWO

Dywergencją $D(y\,||\,z)$, nazywana jest funkcja dwuargumentowa określona na nieujemnych zmiennych *z* i *y,* która spełnia warunek $D(y\,||\,z)\geq 0$, gdzie $D(y\,||\,z)=0$ wtedy i tylko wtedy, gdy $y=z$ [1,5,6]. Dywergencja nie musi natomiast spełniać nierówności trójkąta $D(y\,||\,z)\leq D(y\,||\,x)+D(x\,||\,z)$ oraz nie musi być spełniony warunek symetryczności $D(y\,||\,x)=D(x\,||\,y)$. Dla części dywergencji konieczny jest warunek sumowania się wartości zmiennych z i y do jedności. Dywergencje mogą być zdefiniowane dla wielkości ciągłych jak i dyskretnych. Obecnie funkcje dywergencji stosowane są do oceny podobieństwa (lub jego braku) między nieujemnymi zmiennymi, wektorami, macierzami lub funkcjami. Do najpopularniejszych i najszerszych klas dywergencji zalicza się dywergencję Bregmana oraz dywergencję Csiszar'a [5].

Jedną z popularnych dywergencji jest dywergencji Bosego-Einsteina, która dla wektorów $\mathbf{z}=[z_1,z_2,...,z_L]$ oraz $\mathbf{y}=[y_1,y_2,...,y_L]$, gdzie $y_i,z_i\in[0,1]$ zdefiniowana jest jako

$$D_{BE}^{\alpha}(\mathbf{y}\,\|\,\mathbf{z})=\sum_{i=1}^{L}\left(y_i\ln\frac{(1+\alpha)y_i}{y_i+\alpha z_i}+\alpha z_i\ln\frac{(1+\alpha)z_i}{y_i+\alpha z_i}\right). \tag{2}$$

Dywergencja ta posiada szereg interesujących właściwości, m.in. $D_{BE}^{\alpha}(\mathbf{y}\,\|\,\mathbf{z})=D_{BE}^{1/\alpha}(\mathbf{y}\|\mathbf{z})$ oraz $D_{BE}^{\alpha\to\infty}(\mathbf{y}\,\|\,\mathbf{z})=D_{KL}(\mathbf{y}\|\mathbf{z})$, gdzie D_{KL} oznacza dywergencję Kullback'a-Leiblera.

Istotą cechą dywergencji (2) jest brak symetrii, co może zostać wykorzystane przy ocenie podobieństwa sygnałów. Dla sygnałów, w pewnym ogólnym sensie „statystycznie podobnych" można oczekiwać, że kolejność argumentów nie będzie miała znaczenia. W szczególności, dla przypadku sygnałów losowych nie posiadających z założenia wzorców czy regularności, można założyć, że kolejność argumentów, w niesymetrycznej dywergencji (2) nie ogrywa roli. Oznacza to, że dla sygnałów szumowych v_1,v_2 o tym samym rozkładzie miara dywergencji (2), powinna być symetryczna $D_{BE}(v_1\,\|\,v_2)=D_{BE}(v_2\,\|\,v_1)$. Efekt symetrii, standaryzowanych do przedziału [0,1] sygnałów, może być mierzony jako

$$q=\mathrm{abs}\left(\log\frac{D_{BE}(z\,\|\,v)}{D_{BE}(v\,\|\,z)}\right). \tag{3}$$

Badając symetrię, z wykorzystaniem dywergencji Bosego-Einsteina, należy mieć na uwadze wpływ parametru α na jej wartości. Obecnie zostanie to szerzej przedstawione.

Zwróćmy uwagę, że wyrażenie pod znakiem sumy w (2) można rozpisać jako

$$\begin{aligned}f(y,z)&=y\ln\frac{(1+\alpha)y}{y+\alpha z}+\alpha z\ln\frac{(1+\alpha)z}{y+\alpha z}=\\&=y\ln\big((1+\alpha)y\big)-y\ln\big(y+\alpha z\big)+\alpha z\ln\big((1+\alpha)z\big)-\alpha z\ln\big(y+\alpha z\big)\end{aligned}. \tag{4}$$

Dla $\alpha \in (0,1)$, wszystkie wielkości logarytmowane przyjmują wartości z zakresu $(0,2)$. Dla takiego przypadku, możliwe jest następujące rozwinięcie $\ln(x)$ w szereg Taylora

$$\ln x = (x-1) - \frac{(x-1)^2}{2} + ... + (-1)^{n+1} \frac{(x-1)^n}{n} + ... \approx x - 1 . \tag{5}$$

W efekcie

$$\begin{aligned} & f(y,z) \approx y\big((1+\alpha)y-1\big) - y\big(y+\alpha z-1\big) + \\ & + \alpha z\big((1+\alpha)z-1\big) - \alpha z\big(y+\alpha z-1\big) = \alpha (y-z)^2 . \end{aligned} \tag{6}$$

Uwzględniając (2) i (6) dla $\alpha \in (0,1)$ zachodzi

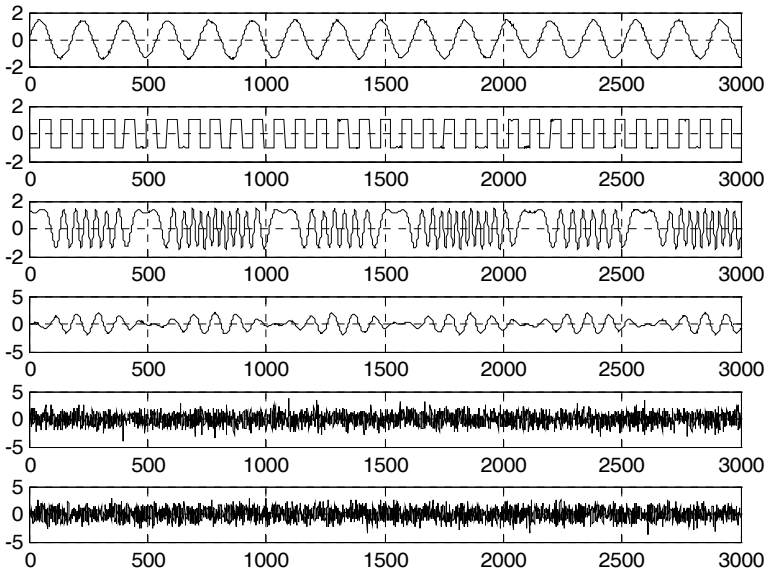
$$D_{BE}^{\alpha}(\mathbf{y} \,||\, \mathbf{z}) = \alpha \left\| \mathbf{y} - \mathbf{x} \right\|_2^2 + R , \tag{7}$$

gdzie *R* oznacza resztę aproksymacji. Przy czym, z (7) wynika, że to *R* odpowiada za efekt asymetrii $D_{BE}^{\alpha}(\mathbf{y} \,||\, \mathbf{z})$. Oznacza to, że od doboru parametru α zależy czułość metody w zakresie oceny podobieństwa. Jednak jak zauważmy w części eksperymentalnej, nawet przyjęcie symetryzującego $\alpha \in (0,1)$, pozwala na osiągnięcie satysfakcjonujących wyników.

EKSPERYMENT PRAKTYCZNY

Powyższe rozważania zostaną obecnie zaprezentowane w kontekście symulowanych badań komputerowych. Postawiony wyżej problem, poszukiwania dobrej, „intuicyjnej", ale jednocześnie ilościowej miary podobieństwa, obrazuje rysunek 1. Widoczne na nim sygnały są wzajemnie niezależne (a więc i zdekorelowane), mają także jednostkowe wariancje.

Rysunek 1. Podobieństwo a korelacja



Źródło: opracowanie własne

Odpowiadająca tym sygnałom macierz współczynników korelacji **C** ma postać

$$\mathbf{C} = \begin{bmatrix} 1.0000 & 0.0594 & -0.0526 & -0.0030 & -0.0502 & -0.0033 \\ 0.0594 & 1.0000 & -0.0098 & 0.0098 & 0.0420 & 0.0142 \\ -0.0526 & -0.0098 & 1.0000 & 0.0226 & -0.0460 & -0.0087 \\ -0.0030 & 0.0098 & 0.0226 & 1.0000 & -0.0108 & -0.0204 \\ -0.0502 & 0.0420 & -0.0460 & -0.0108 & 1.0000 & -0.0084 \\ -0.0033 & 0.0142 & -0.0087 & -0.0204 & -0.0084 & 1.0000 \end{bmatrix}$$

Widać, że korelacyjna ocena podobieństwa, nie odpowiada wizualnej ocenie, w szczególności w odniesieniu do dwóch ostatnich szumowych sygnałów. Z kolei, zastosowanie odległości opartej na p-normie przy p=2, prowadzi do wyników prezentowanych w Tabeli 1.

Tabela 1. Odległości między sygnałami mierzone p-normą, przy p=2

	S1	S2	S3	S4	S5	S6
S1	0	0,0333	0,0409	0,0316	0,0287	0,0287
S2	0,0333	0	0,0379	0,0267	0,0226	0,0239
S3	0,0409	0,0379	0	0,0351	0,0331	0,0335
S4	0,0316	0,0267	0,0351	0	0,0184	0,0197
S5	0,0287	0,0226	0,0331	0,0184	0	0,0153
S6	0,0287	0,0239	0,0335	0,0197	0,0153	0

Źródło: opracowanie własne

Jak widać, choć podobieństwo między sygnałami jest stosunkowo właściwie oceniane, to jednocześnie symetryczność tej oceny, istotnie ogranicza dalszą interpretację. Nie sposób ocenić także, jaki charakter mają analizowane sygnały. Pomiar odległości za pomocą dywergencji Bosego-Einsteina nawet przy symetryzującym parametrze $\alpha = 0{,}5$ pozwala na wyraźne odróżnienie sygnałów podobnych, co prezentuje Tabela 2.

Tabela 2. Odległości między sygnałami mierzone symetryzowaną dywergencją Bosego-Einsteina przy $\alpha = 0{,}5$

	S1	S2	S3	S4	S5	S6
S1	0	0,2354	0,4102	0,2038	0,1661	0,1674
S2	0,2415	0	0,3855	0,1525	0,1113	0,1241
S3	0,375	0,3406	0	0,2978	0,2756	0,2809
S4	0,2197	0,1641	0,3541	0	0,0663	0,0764
S5	0,1876	0,1264	0,3398	0,071	0	0,0435
S6	0,1881	0,1392	0,3437	0,0812	0,0428	0

Źródło: opracowanie własne

Dodatkowo stopień symetryczności oceny podobieństwa za pomocą dywergencji Bosego-Einsteina, pozwala na wnioskowanie o statystycznym podobieństwie typowym dla sygnałów losowych. Wyniki oceny symetrii, mierzonej parametrem (3) prezentuje Tabela 3.

Tabela 3. Stopień symetryczności dla odległości między sygnałami mierzony symetryzowaną dywergencją Bosego-Einsteina przy $\alpha = 0{,}5$

	S1	S2	S3	S4	S5	S6
S1	0	0,0255	0,0898	0,0751	0,1216	0,1165
S2	0	0	0,1237	0,0738	0,1272	0,1147
S3	0	0	0	0,1731	0,2095	0,202
S4	0	0	0	0	0,0684	0,0606
S5	0	0	0	0	0	0,0158
S6	0	0	0	0	0	0

Źródło: opracowanie własne

Otrzymane wyniki potwierdzają zasadność wykorzystania opisanej metody oceny podobieństwa opartej na dywergencjach Bosego-Einsteina. Mogą mięć one szerokie zastosowanie w automatycznych systemach transakcyjnych, w których jednym z podstawowych problemów jest poszukiwanie podobieństw między historycznymi wzorcami. Ma to szczególne znaczenie, gdy do systemów automatycznych chcemy przekazać ludzką wiedzę i doświadczenie. Jak widać z przedstawionego przykładu brak korelacji między danymi nie musi oznaczać braku podobieństw w potocznym tego słowa rozumieniu.

ZAKOŃCZENIE

W artykule przedstawiono koncepcję oceny podobieństwa wykorzystującą niesymetryczne właściwości miar dywergencji. Zastosowana dywergencja Bosego-Einsteina, jako typowy przykład niesymetrycznej dywergencji, może być zastąpiona innym typem dywergencji. Dywergencja Bosego-Einsteina ma interesujące właściwości zmiany stopnia symetryczności w zależności od parametru aplha. Jego zmiany pozwalają na różnicowanie stopnia asymetrii miary, co pozwala dostosować czułość metody do posiadanych danych. Należy także zauważyć, że fakt zakładanej nieujemności sygnałów, nie jest tu zasadniczym ograniczeniem, gdyż oceniany i porównywany jest kształt rzeczywistych sekwencji danych. Dlatego dla danego sygnału jak również szumu referencyjnego, zawsze możliwym jest dokonanie przesunięcia lub standaryzacji do wartości nieujemnych.

BIBLIOGRAFIA

Amari, S. (1985) Diferential-Geometrical Methods in Statistics. Springer Verlag.

Anscombe, F. J. (1973) Graphs in statistical analysis. The American Statistician 27: 17–21.

Cardoso J.-F. and Comon P. (1996): Independent component analysis, a survey of some algebraic methods. In Proc. ISCAS Conference, volume 2, pages 93–96, Atlanta.

Cichocki, A., Zdunek, R., Amari, S.: (2006) Csiszar's Divergences for Non-Negative Matrix Factorization: Family of New Algorithms, Lecture Notes in Computer Science, Vol. 3889, pp. 32--39, Springer Verlag, Heidelberg.

Cichocki, A., Zdunek, R., Phan, A.-H., Amari, S. (2009) Nonnegative Matrix and Tensor Factorizations: Applications to Exploratory Multi-way Data Analysis. John Wiley.

Csiszar, I. (1974) Information measures: A critical survey. In: Prague Conference on Information Theory. Volume A, pp. 73--86. Academia Prague.

Krutsinger J. (1997) Trading Systems: Secrets of the Masters, McGraw-Hill.

Luo, Y., Davis, D., Liu, K.: (2002) A Multi-Agent Decision Support System for Stock Trading. In: The IEEE Network Magazine Special Issue on Enterprise Networking and Services, vol.16, No. 1.

Rodgers J. L. and Nicewander W. A. (1988) Thirteen ways to look at the correlation coefficient. The American Statistician, 42(1):59–66.

Therrien, C.W. (1992) Discrete Random Signals and Statistical Signal Processing. Prentice Hall, New Jersey.

BOSE-EINSTEIN DIVERGENCES FOR SIMILARITY ANALYSIS IN FINANCIAL TIME SERIES

Abstract: The similarity assessment of the financial time series is the one of problems where the proper methodological choice is very important. The typical correlation approach can lead to misleading results. Often the similarity score is contrary to the visual observations, expert's knowledge and even a common sense. The reasons of such situations can be associated with the properties of the correlation measure and its adequateness for analyzed data, as well as in terms of methodology aspects. In this article, we point these disadvantages associated with the use of correlation to assess the similarity of financial time series as well as we propose the alternative solution based on divergence measures. In particular, we focus on the Bose-Einstein divergence. The practical experiments with simulated data confirm the validity of our concept.

Keywords: time series similarity, divergence measures, Bose-Einstein divergence