Folia Oeconomica
Acta Universitatis Lodziensis
ISSN 0208-6018 e-ISSN 2353-7663
www.czasopisma.uni.lodz.pl/foe/
5(338) 2018

DOI: http://dx.doi.org/10.18778/0208-6018.338.01

Michał Bernardelli
Warsaw School of Economics, College of Economic Analysis, Institute of Econometrics,
mbernard@sgh.waw.pl

Hidden Markov Models as a Tool for the Assessment of Dependence of Phenomena of Economic Nature[1]

Abstract: The assessment of dependence between time series is a common dilemma, which is often solved by the use of the Pearson's correlation coefficient. Unfortunately, sometimes, the results may be highly misleading. In this paper, an alternative measure is presented. It is based on hidden Markov models and Viterbi paths. The proposed method is in no way universal but seems to provide quite an accurate image of the similarities between time series, by disclosing the periods of convergence and divergence. The usefulness of this new measure is verified by specially crafted examples and real-life macroeconomic data. There are some definite advantages to this method: the weak assumptions of applicability, ease of interpretation of the results, possibility of easy generalization, and high effectiveness in assessing the dependence of different time series of an economic nature. It should not be treated as a substitute for the Pearson's correlation, but rather as a complementary method of dependence measure.

Keywords: dependence measure, correlation, hidden Markov model, Viterbi path

JEL: C63, E24, C18

[1] The views and opinions expressed in this paper are those of the authors and do not necessarily reflect the views and opinions of the National Bank of Poland. The project entitled "Discussion Forum – Measurement and Evaluation of Economic and Social Phenomena" (MASEP2017) is implemented in cooperation with the National Bank of Poland within the framework of economic education.

1. Introduction

The concept of measuring dependence between variables is, without a doubt, one of the most important problems in modern theory. Similarity measures are also crucial parts of many applications in time series data mining, e.g. clustering and classification. Over the years several approaches have been proposed. Despite the existence of methods dedicated to time series analysis, like cointegration or copulas, the most known and used method of accessing the relationship between two time series is definitely Pearson's correlation coefficient, which, in fact, has a much more general use. However, there are well known theoretical cases, where correlation is not a good measure of dependence. It turns out, that in economic practice exploring correlation may be quite misleading.

The purpose of this paper is to present an alternative to the existing measures of dependence between time series. This measure, in some cases, seems to be a better reflection of dependence when compared to the Pearson's correlation coefficient. It also requires much weaker assumptions than common econometric or statistic methods for time series analysis. The proposed method employs hidden Markov models (HMM) and Viterbi paths. These tools are widely used in the areas, where the pattern recognition is explored. They may also be used to find similarities between time series, and therefore, in some cases, may give more accurate results than the Pearson's correlation coefficient. This may be especially useful for non-linear data. What is more, this approach allows for the specification of periods of convergence and divergence between the data. The effectiveness of the method was verified based on specially prepared test data, but most of all, based on the exemplary data from the Central Statistical Office of Poland. There is, however, no comparison of effectiveness given between the proposed method and other existing approaches dedicated to time series relationship analysis. The main goal of this paper is to give a detailed description of the method and to present it as a comparison to the most popular measure, that is the Pearson's correlation coefficient.

This paper consists of 6 sections. After the introduction, in section 2, some common dependence measures are briefly presented, with the emphasis on pros and cons of the Pearson's correlation coefficient. Section 3 discusses the key tools in the formulation of the new dependence measure, that is: the hidden Markov models and Viterbi paths. Section 4 contains a detailed description of the new dependence measure, which is complemented by examples in section 5. The article ends with the conclusion in Section 6.

2. Time series similarity measures

Various ways of assessing the dependence exist and are used in data analysis. There is no simple classification of available methods, but some effort has been made to unify the comparison methods (Parzen, Mukhopadhyay, 2012). Some of the measures, such as correlation coefficients, have been well known for over a half of a century (Kendall, Stuart, 1973; Soper et al., 1917), some, such as distance correlation (Székely, Rizzo, Bakirov, 2007) or local Gaussian correlation function (Tjostheim, Hufthammer, 2013), are relatively new. Many time series applications are related to the similarity search and exploit methods such as discrete Fourier transform or wavelet transform (Wu, Agrawal, Abbadi, 2000). The best known similarity measures dedicated only to time series, are the cointegration method and copulas. A comprehensive introduction to copula theory and dependence modeling can be found in books of Joe (1997) and Nelsen (2006). The theory and discussion of the time series cointegration can be found in Dhrymes (1997) or Maddala and Kim (1998).

An extensive comparison of similarity measures for time series classification can be found in Lhermitte et al. (2011) or Serrà and Arcos (2014). The comparisons include similarity measures such as distance measures (Euclidean, Manhattan, Mahalanobis), correlation based measures, dynamic time warping, Fourier based similarities, and principal component analysis.

Nevertheless, the best-known method of measuring the dependence is definitely a classic Pearson's correlation. It was introduced by Francis Galton and Karl Pearson (1895) at the end of the 19th century. It is a simple measure of the linear correlation between two variables, denoted usually[2] by the letter r, and given by the formula for the n-element dataset $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$,

$$r = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}, \tag{1}$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and are $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ the sample means.

The Pearson's correlation coefficient ranges from −1 to 1. Values –1 and 1 suggest a perfect linear relationship, whereas values close to zero imply that there is no linear relationship between the variables. There exist some arbitrary rules or guidelines for the interpretation of a correlation coefficient (Guilford, 1956), however, the interpretation always depends on the context and can't be generalized to all applications.

[2] Sometimes the name sample Pearson correlation coefficient is used. For the populations, Pearson's correlation coefficient is represented by the letter ρ.

Although the Pearson's correlation coefficient is widely used in the sciences, it has some drawbacks. Most importantly, it was developed only to assess the degree of linear relationship. One of the other disadvantages is its sensitivity to outliers. For example, in Figure 1, there are two time series with perfect positive correlation (parallel lines) at all times except the last three points. Pearson's correlation coefficient for these sample datasets equals 0.6951. The interpretation could imply that the two corresponding datasets are not so similar after all, whereas there are only 3 out of 50 points causing discrepancies.

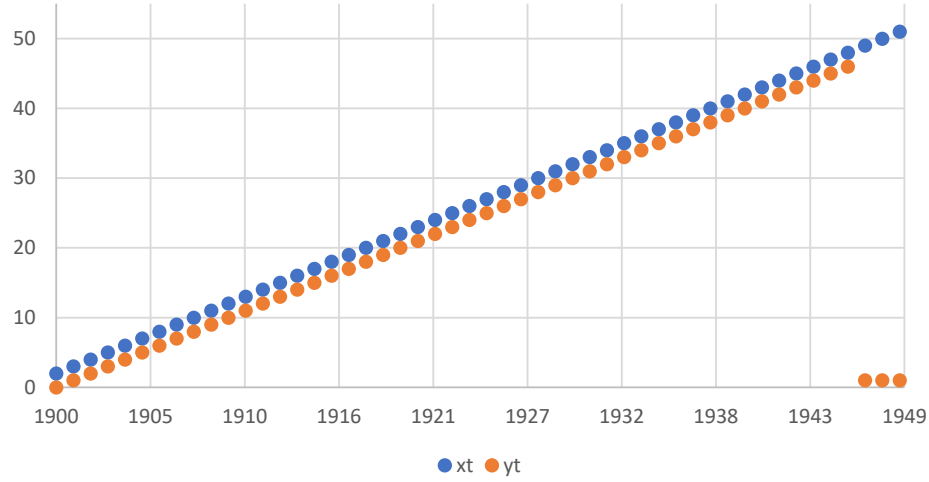

Figure 1. Example of the influence of outliers on the value of the Person's correlation coefficient. For exemplary datasets $r = 0.6951$

Source: own calculations

Having considered presented arguments and examples, one must agree that correlation, Pearson's in particular, is not a good measure of dependence in some cases. In economic practice, drawing conclusions based on this correlation coefficient may be highly misleading. Therefore, it is important to develop alternative dependence measures, which is the main goal of this paper. It must be emphasized that the Pearson's correlation coefficient is designed to measure the correlation between the variables, not the time series. One of the key features is the immutability for the permutation of elements. In contrast, the method proposed in this paper is dedicated to the time series only.

3. Hidden Markov models and Viterbi path

In this section, a brief introduction to the theory of hidden Markov models (HMM) is presented. Combined with the concept of a Viterbi path, HMMs are the basis of the new measure, described in the next section.

The HMMs are present in the literature at least since the 1960s of the previous century but are usually identified with the name of Hamilton (1989). Hidden Markov models are the generalization of the Markov models (Cappé, Moulines, Rydén, 2005) achieved by an addition of an extra layer. Formally, HMM $\{X_k, Y_k\}_{k\geq 0}$ is a discrete stochastic process satisfying the following conditions:

1) the unobservable process $\{X_t\}_{t\geq 0}$ is a homogenous MC with a finite state space S,
2) conditionally on the process $\{X_t\}_{t\geq 0}$ the observations $\{Y_t\}_{t\geq 0}$ are independent, and for each t the conditional distribution of Y_t depends only on X_t.

In macroeconomic applications, the normal HMM is often used, which refers to the case where Y_t has a Gaussian distribution. HMM are widely used in the areas, where the pattern recognition is explored, such as speech, handwriting, gesture or voice recognition. HMM is also used in bioinformatics (e.g. DNA sequencing process) or macroeconomics (e.g. business cycles synchronization analysis, turning points identification).

In HMM the states are unobservable, and a few algorithms for calculating them exist. All of them are based on another observable time series, to be more precise, on the estimated transition probabilities and the parameters of the probability distribution related to each state. Estimation of the HMM parameters may be done with the use of the Baum-Welch algorithm (Baum et al., 1970), whereas to find the most probable path of states, the concept of smoothed or filtered probabilities can be exploited. Sometimes, the path of states may be optimal only locally, therefore it is advisable to use a more effective approach called Viterbi algorithm (Viterbi, 1967), which takes under consideration the whole period covered by the analysis. To be more formal, the Viterbi path is the path of states $(x_1^*, x_2^*, \dots, x_T^*) \in S^T$ such, that

$$P(X_1 = x_1^*, X_2 = x_2^*, \dots, X_T = x_T^* | Y_1 = y_1, Y_2 = y_2, .., Y_T = y_T) = \max_{(x_1, x_2, \dots, x_T) \in S^T} \{P(X_1 = x_1, X_2 = x_2, \dots, X_T = x_T | Y_1 = y_1, Y_2 = y_2, \dots, Y_T = y_T)\}. \quad (2)$$

As a basis of the new dependence measure, presented in the next section, the Viterbi algorithm together with the Baum-Welch algorithm were used. Both of these algorithms are completely deterministic but their results strongly depend on the initial values and can be far from optimal. To increase the chance of finding the globally optimal solution, the Monte Carlo simulations were used (Bernardelli, 2013). In computations, 10000 repetitions were performed, because the present-

ed examples proved to be stable. We restricted ourselves to the analysis of normal HMM with two-element state space $S = \{0, 1\}$, where time series under the analysis must satisfy the conditions

$$Y_n|_{X_n=0} \sim N(\mu_0, \sigma_0) \text{ and } Y_n|_{X_n=1} \sim N(\mu_1, \sigma_1). \tag{3}$$

We additionally assume that $\mu_0 < \mu_1$ to have the same order of states in each considered case (state 1 is associated with a greater mean value).

The presented procedure, involving Monte Carlo simulations, Baum-Welch algorithm, and Viterbi algorithm, allows for the finding of the optimal path of states for the considered time series. This path, called the Viterbi path, is the foundation of the new dependence measure.

4. Dependence measure based on hidden Markov model

In the previous section, the concept of HMM and Viterbi path was described. In this section the description of the measure for assessing the relationship between time series is presented. The procedure, in order to get the result in the empirical analysis, explores the concept of HMM and Viterbi path described in the previous section.

The procedure of calculation can be described in the following steps:

1. Normalization of time series x_t and y_t

$$\widetilde{x_t} = \frac{x_t - \min\limits_s x_s}{\max\limits_\tau \left| x_\tau - \min\limits_s x_s \right|} \text{ and } \widetilde{y_t} = \frac{y_t - \min\limits_s y_s}{\max\limits_\tau \left| y_\tau - \min\limits_s y_s \right|} \tag{4}$$

   This step is necessary because time series can be expressed in different units and sizes. After normalization $\widetilde{x_t}, \widetilde{y_t} \in [0; 1]$. Of course, there are many other methods of normalizatvion (Walesiak, 2016), however, formula (4) gives the values suitable for an input to the HMM.

2. Calculation of the difference between normalized time series. Depending on the sign of the Pearson's correlation coefficient we define

$$\widetilde{z_t} = \frac{(\widetilde{x_t} - \widetilde{y_t}) - \min\limits_s (\widetilde{x_s} - \widetilde{y_s})}{\max\limits_\tau \left| (\widetilde{x_\tau} - \widetilde{y_\tau}) - \min\limits_s (\widetilde{x_s} - \widetilde{y_s}) \right|} \tag{5}$$

for positively correlated time series $\widetilde{x_t}$ and $\widetilde{y_t}$, and

$$\widetilde{z_t} = \frac{(\widetilde{x_t}+\widetilde{y_t}) - \min_s(\widetilde{x_s}+\widetilde{y_s})}{\max_\tau\left|(\widetilde{x_\tau}+\widetilde{y_\tau}) - \min_s(\widetilde{x_s}+\widetilde{y_s})\right|} \tag{6}$$

for negatively correlated time series $\widetilde{x_t}$ and $\widetilde{y_t}$. Formulas (5) and (6) are, in fact, the same normalization as in (4) but for time series $\widetilde{x_t} - \widetilde{y_t}$ or $\widetilde{x_t} + \widetilde{y_t}$, depending on the direction of the Pearson's correlation.

3. Calculation of the parameters of HMM and Viterbi path for time series $\widetilde{z_t}$ constructed in step 2. The procedure was described in the previous section. Let v_t denote the Viterbi path for $\widetilde{z_t}$. States 0 on this path represent periods where original time series x_t and y_t are similar and states 1 may be interpreted as a period in which given time series diverge.
4. As a resulting value of the new measure, average number of states 0 on the Viterbi path v_t is proposed. We will denote this number as

$$r_{HMM} = \frac{numer\ of\ states\ 0\ on\ v_t}{length\ of\ v_t} \in [0;1]. \tag{7}$$

The intuition behind r_{HMM} is simple: calculate what percentage of time the two given variables are similar and compare that to the length of the whole considered period. For ideal convergence $r_{HMM} = 1$, and for complete divergence $r_{HMM} = 0$.

The idea should be clearer after the steps of the procedure for the example given in Figure 1 have been performed. Step 1 is omitted in the presentation because in this simple case the figure with normalized time series is very similar to the figure with the original time series (of course the scale on y-axis will differ). In Figure 2 the time series z_t (top) and v_t (bottom) are presented.

The difference between time series for most of the period (except the last three points) is close to zero, therefore corresponding states on the Viterbi path are the zero states. At the end of the considered time, the difference of the normalized time series increases dramatically, and at the same time (time-points with indexes 48, 49 and 50) the states of the Viterbi path change from 0 to 1. The last step is the final calculation of $r_{HMM} = \frac{47}{50} = 0.94$. Compare this result with Pearson's correlation coefficient equal 0.6951. Looking at the original time series, the new measure definitely describes dependence more accurately than the Pearson's correlation. After all, the time series are similar exactly 94% of the time. Of course, the still unresolved issue is the comparison to other methods, especially these designed for the time series only. We will not give this kind of comparison. We will, however, state some facts about the advantages of the proposed method. One of them is the lack of econometric character assumptions. The second, is the ease of the interpretation of the results. And the last but not least is the significant advantage of the possibility of phases identification, when the time series are similar and the times when they are not.

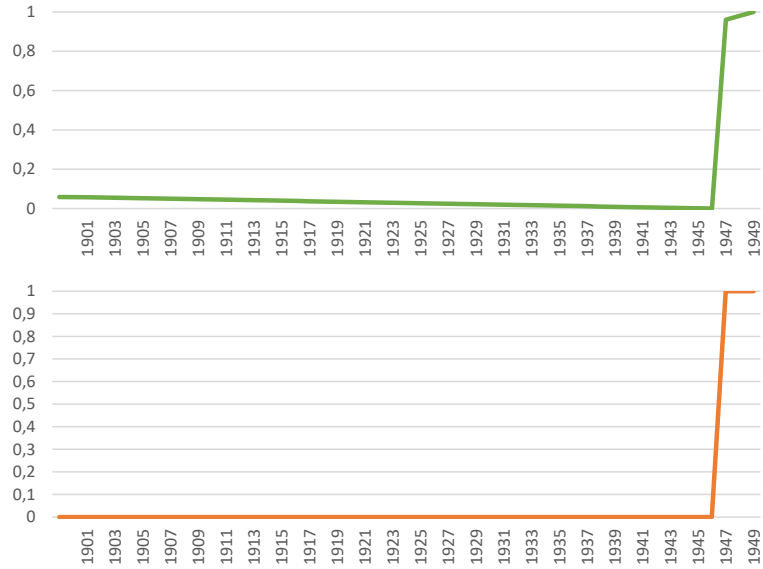

Figure 2. Time series z_t from step 2 (top figure) and corresponding Viterbi path v_t (bottom figure)
Source: own calculations

In the next section other examples and illustration of the comparison of the proposed dependence measure with the Pearson's correlation coefficient are presented. Among these examples, there are also time series from the Central Statistical Office of Poland.

5. Real-life examples

Some examples are given in this section to illustrate the usefulness of the dependence measure based on the hidden Markov models. In the example given in the previous section the value r_{HMM} was much greater than the value of the Pearson's correlation coefficient r. The next example is also artificial, and illustrates the opposite relation between two measures. The example consists of two shifted sine functions (see Figure 3). Time series after normalization are given in Figure 4. The shape of the normalized time series and the original ones are similar, but the values are different: [–1; 1] vs. [0; 1] (see the y-axis). The graphs of the time series from the second (z_t) and third (v_t) steps of the procedure are given in Figure 5. The periods of convergence and divergence are easily visible. Calculation of the percentage of the 0 states in the Viterbi path results in the value of the measure $r_{HMM} = 0.56$, which is smaller compared to the Pearson's correlation coefficient $r = 0.87$. The shift between the original time series is big enough to imply similarities about 56% of the time. Choosing a smaller shift we would get a greater value of coefficient r_{HMM}.

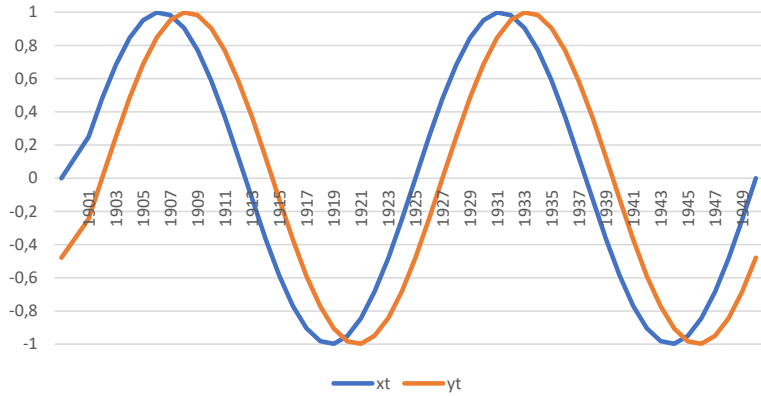

Figure 3. Two sine functions (x_t and y_t) shifted relative to each other

Source: own calculations

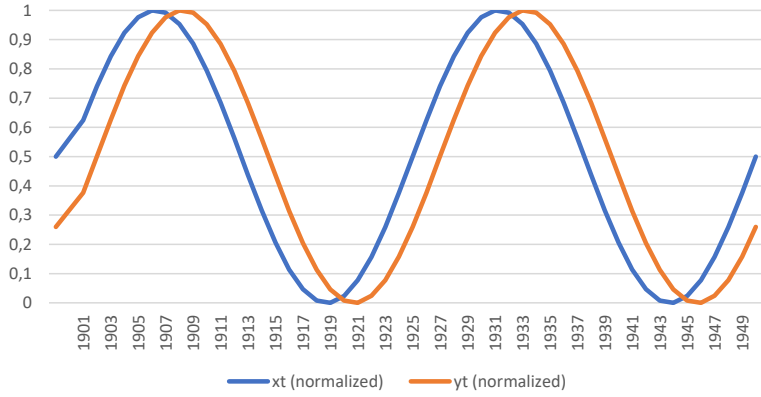

Figure 4. Two sine functions (and) after normalization (step 1)

Source: own calculations

The last two examples are based on the data from the Central Statistical Office of Poland. In the first one we are interested in checking the dependence between the number of working women and men in Poland in years 1995–2016. Time series before and after normalization are given in Figure 6.

Before normalization time series seem to be similar at the end of the considered period. After normalization, we can see that the behavior of the time series is alike rather in the middle period. According to the Viterbi path (see Figure 7), a number of working men and women in Poland were correlated until 2008, and after that year we can talk about the divergence between the number of people in the considered groups. The resulting value of the measure $r_{HMM} = 0.64$ indicates much smaller relationship than the Pearson's correlation coefficient $r = 0.84$. The verification of which out of the two measures is more accurate is, of course, arbitrary, but thinking about

economic interpretation, it can be clearly seen, that the number of people in the two groups of men and women changes at a different pace, starting around 2009. Keeping this in mind, the measure proposed in this article seems to be better suited.

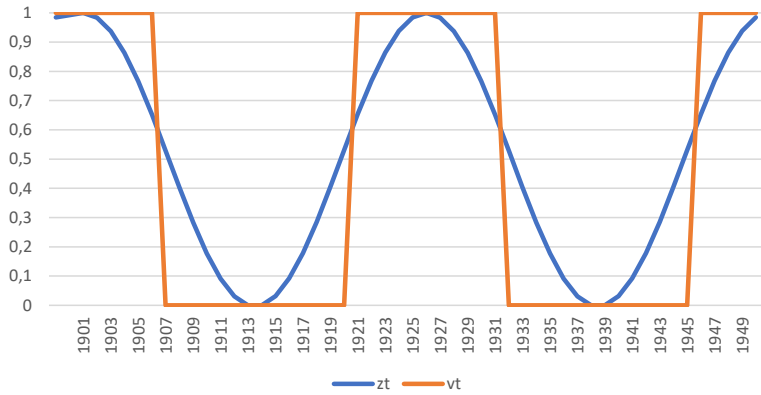

Figure 5. Time series z_t (solid line) and Viterbi path v_t (dashed line) for the time series from Figure 3

Source: own calculations

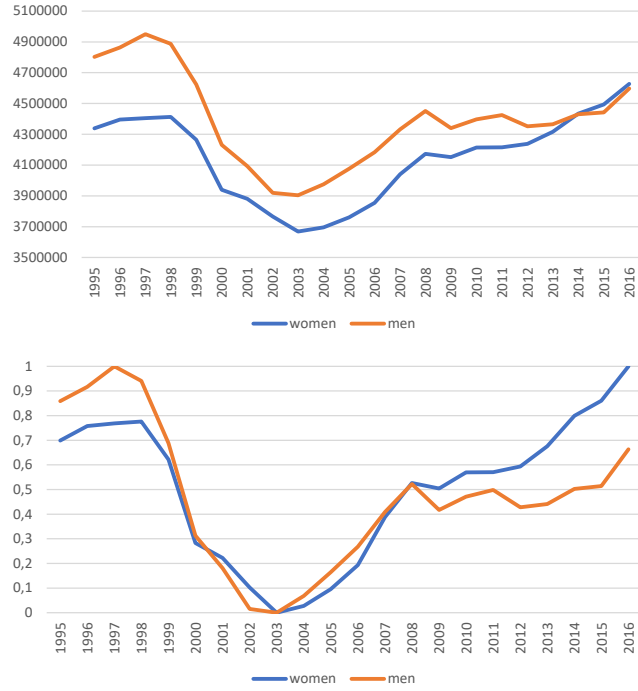

Figure 6. Number of working people by sex in Poland in 1995–2016 before (top) and after (bottom) normalization

Source: own calculations

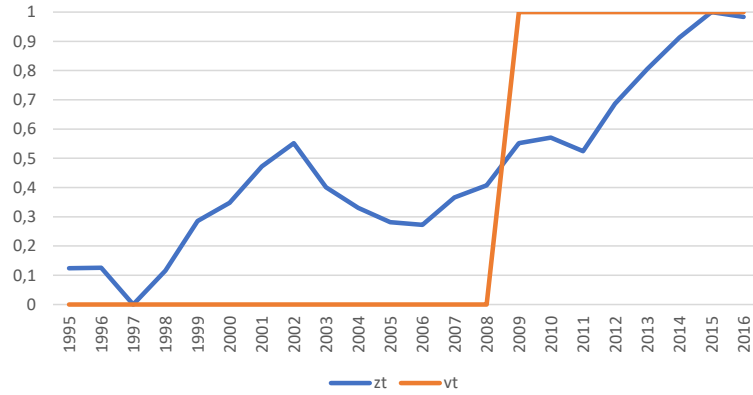

Figure 7. Time series z_t from step 2 (solid line) and a corresponding Viterbi path v_t (dashed line) for the time series from Figure 6

Source: own calculations

In the last example a number of marriages in two neighboring voivodships: lodzkie and mazowieckie, are compared. The time series after normalization are given in Figure 8.

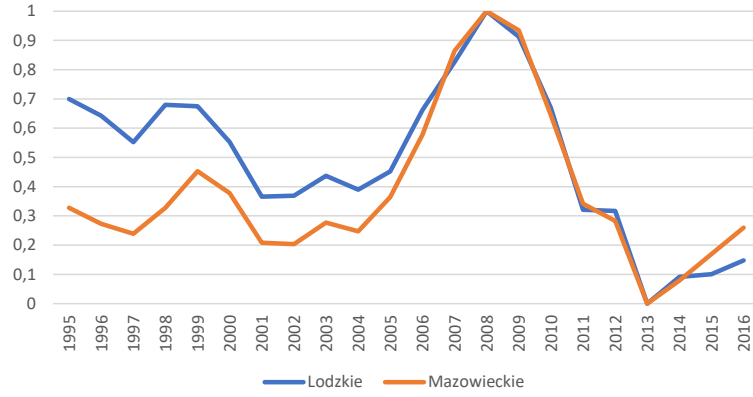

Figure 8. A number of marriages in lodzkie and mazowieckie in 1995–2016 (after normalization)

Source: own calculations

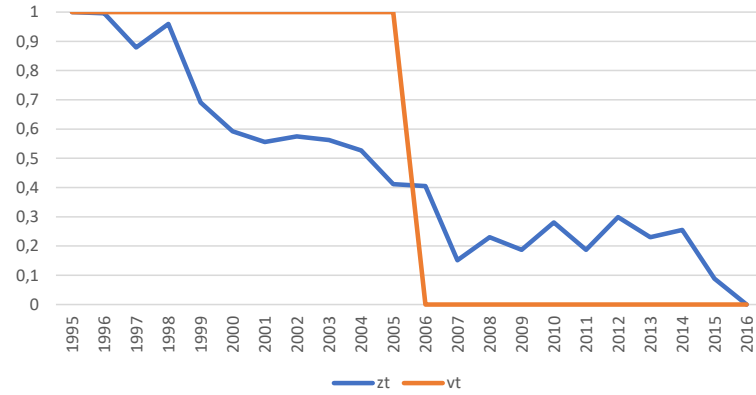

Figure 9. Time series z_t from step 2 (solid line) and a corresponding Viterbi path v_t (dashed line) for the time series from Figure 8

Source: own calculations

The Viterbi path for this example is presented in Figure 9. The values of the correlation coefficient are as follows $r_{HMM} = 0.85$ and $r = 0.85$. As before, it seems that the measure based on HMM gives a more accurate assessment of the dependence, than the Pearson's correlation coefficient.

6. Conclusions

HMM proved to be an effective method of analyzing the macroeconomic time series in many applications. It was already used in turning point identification, synchronization of the business cycles, and analysis of convergence. In this article, the new measure of dependence between time series involving the use of the HMM and the Viterbi path was shown. The advantages of this method are mostly: the weak assumptions of applicability and the ease of interpretation of the results. In the given measure only two state HMM was used, but it is probably worth noticing, that the generalization to more than two states is possible. Examples from the previous sections should provide the evidence for the usefulness of this measure in the analysis of the time series of the economic character. The measure itself, however, should not be treated as a substitute for the Pearson's correlation, but rather as a complementary method to the dependence measure.

References

Baum L.E., Petrie T., Soules G., Weiss N. (1870), *A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains*, "The Annals of Mathematical Statistics", vol. 41, no. 1, pp. 164–171.

Bernardelli M. (2013), *Non-classical Markov Models in the Analysis of Business Cycles in Poland*, "Annals the Collegium of Economic Analysis", vol. 30, pp. 59–74.

Cappé O., Moulines E., Rydén T. (2005), *Inference in Hidden Markov Models*, Springer Series in Statistics, Springer-Verlag, New York.

Dhrymes P.J. (1997), *Time Series, Unit Roots, and Cointegration*, Academic Press, San Diego.

Guilford J.P. (1956), *Fundamental statistics in psychology and education*, McGraw-Hill, New York.

Hamilton J.D. (1989), *A New Approach to the Economic Analysis of Non-stationary Time Series and Business Cycle*, "Econometrica", no. 57, pp. 357–384.

Joe H. (1997), *Multivariate Models and Dependence Concepts*, Monographs in Statistics and Applied Probability (Book 73), Chapman and Hall, London.

Kendall M.G., Stuart A. (1973), *The Advanced Theory of Statistics*, vol. 2: *Inference and Relationship*, Griffin, New York.

Lhermitte S., Verbesselt J., Verstraeten W.W., Coppin P. (2011), *A comparison of time series similarity measures for classification and change detection of ecosystem dynamics*, "Remote Sensing of Environment", vol. 115(12), pp. 3129–3152.

Maddala G.S., Kim I. (1998), *Unit Roots, Cointegration, and Structural Change*, Cambridge University Press, Cambridge, pp. 155–248.

Nelsen R.B. (2006), *An Introduction to Copulas*, Second Edition, Springer-Verlag New York.

Parzen E., Mukhopadhyay S. (2012), *Modeling, dependence, classification, united statistical science, many cultures*, https://arxiv.org/abs/1204.4699 [accessed: 20.01.2018].

Pearson K. (1895), *Notes on regression and inheritance in the case of two parents*, "Proceedings of the Royal Society of London", vol. 58, pp. 240–242.

Serrà J., Arcos J.L. (2014), *An Empirical Evaluation of Similarity Measures for Time Series Classification*, Knowledge-Based Systems, vol. 67, pp. 305–314.

Soper H.E., Young A.W., Cave B.M., Lee A., Pearson K. (1917), *On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R.A. Fisher. A co-operative study*, "Biometrika", vol. 11, pp. 328–413.

Székely G.J., Rizzo M.L., Bakirov N.K. (2007), *Measuring and testing dependence by correlation of distances*, "The Annals of Statistics", vol. 35, no. 6, pp. 2769–2794.

Tjostheim D., Hufthammer K.O. (2013), *Local Gaussian correlation: A new measure of dependence*, "Journal of Econometrics", vol. 172, issue 1, pp. 33–48.

Viterbi A. (1967), *Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm*, "IEEE Transactions on Information Theory", vol. 13, pp. 260–269.

Walesiak M. (2016), *The choice of groups of variable normalization methods in multidimensional scaling*, "Przegląd Statystyczny", R. LXIII, no. 1, pp. 7–18.

Wu Y., Agrawal D., Abbadi A.E. (2000), *A comparison of DFT and DWT based similarity search in time-series databases*, Proceedings of the 9th International Conference on Information and Knowledge Management, McLean.

Ukryte modele Markowa jako narzędzie oceny zależności zjawisk o charakterze ekonomicznym

Streszczenie: Ocena zależności między szeregami czasowymi jest zagadnieniem, które jest często rozwiązywane za pomocą współczynnika korelacji Pearsona. Niestety, czasami wyniki mogą być bardzo mylące. W artykule przedstawiono alternatywną miarę badania zależności, opartą na ukrytych modelach Markowa oraz ścieżkach Viterbiego. Zaproponowana metoda nie jest uniwersalna, ale wydaje się dość dokładnie odzwierciedlać podobieństwo między szeregami czasowymi, eksponując okresy zbieżności i rozbieżności. Przydatność tej nowej miary została zweryfikowana na przykładach, jak również realnych danych makroekonomicznych. Zaletami tej metody są: słabe założenia stosowalności, łatwość interpretacji wyników, możliwość generalizacji i wysoka skuteczność w ocenie zależności różnych szeregów czasowych o charakterze ekonomicznym. Nie należy jej jednak traktować jako substytutu korelacji Pearsona, a raczej jako uzupełniającą metodę pomiaru zależności.

Słowa kluczowe: miara zależności, korelacja, ukryty model Markowa, ścieżka Viterbiego

JEL: C63, E24, C18

Received: 2017-10-15; verified: 2018-03-06. Accepted: 2018-07-30