Motrenko A.P., Strijov V.V. Extracting fundamental periods to segment human motion time series // Journal of Biomedical and Health Informatics, 2015, 20(6), 1466 - 1476.

Extracting fundamental periods to segment biomedical signals

Anastasia Motrenko, Vadim Strijov,

***Abstract*—We address the problem of segmenting nearly periodic time series into period-like segments. We introduce a definition of nearly periodic time series via triplets ⟨basic shape, shape transformation, time scaling⟩ that covers a wide range of time series. To split the time series into periods we select a pair of principal components of the Hankel matrix. We then cut the trajectory of the selected principal components by its symmetry axis, thus obtaining half-periods that are merged into segments. We describe a method of automatic selection of periodic pairs of principal components, corresponding to the fundamental periodicity.**

We demonstrate the application of the proposed method to the problem of period extraction for accelerometric time series of human gait. We see the automatic segmentation into periods as a problem of major importance for human activity recognition problem, since it allows to obtain interpretable segments: each extracted period can be seen as an ultimate entity of gait.

The method we propose is more general compared to the application specific methods and can be used for any nearly periodical time series. We compare its performance to classical mathematical methods of period extraction and find that it is not only comparable to the alternatives, but in some cases performs better.

***Index Terms*—sensor signal processing, nearly periodic time series, time series segmentation, period extraction, principal components analysis.**

I. Introduction

RECENT advances in wireless technologies make it possible to obtain significant amounts of human-driven data with the help of various wearable devices by forming sensor networks to monitor patient's state at any time. The analysis of sensor networks data allows to solve problems involved in such health applications as analysis of human behavior and social interactions [1], [2], emotion recognition [3] recognition of depressive and manic states and detect state changes of patients suffering from bipolar disorder [4], automated fall detection for elderly people [5] and others [6], [7]. Dealing with human-driven data, one frequently encounters nearly periodic signals, for example when analysing brain's electrical activity [8], pulse wave, heartbeat [9], breathing rate [10] or basic types of human gait [11]. The problem of partitioning a nearly periodic time series into period-like segments is an important part of biosignal analysis and can be used in many biomedical and health applications.

A. Motrenko is with the Department of Applied Mathematics and Control, Moscow Institute of Physics and Technology, Moscow, 141700 Russian Federation e-mail: anastasia.motrenko@gmail.com.

V. Strijov is with the Department of Applied Mathematics and Control, Moscow Institute of Physics and Technology, Moscow, 141700 Russian Federation and Dorodnicyn Computing Centre of RAS, Moscow, 119333 Russian Federation e-mail: strijov@ccas.ru.

The method we propose is applicable to any nearly periodic time series, which enables automatic interpretable segmentation of various biosignals. We see the problem of interpretable segmentation as a step of hierarchical analysis of complex sequences of variously structured time series. In case of accelerometry data [12], the lowest levels of this hierarchy correspond to segments of time series, followed by time series of particular activity types and the highest levels describe sequences of activities. The proposed method of period extraction of human gait time series provides interpretability of time series segmentation and is, therefore, an important step in understanding and modelling human.

There have been developed plenty of highly accurate methods of period extraction, which rely on application specific information, such as the expected form of a signal or average frequency range. Such *ad hoc methods* are specifically designed to recognize certain types of events, such as the acceleration peaks at heel landing and take-off in step detection in case of gait segmentation. Performance of these methods depends on a range of parameters that require accurate tuning. This, on one hand, might be the strength of ad hoc methods which ensures their accuracy (provided this tuning is feasible), but might turn out as their deficiency when such tuning can be afforded. For such cases we develop a method of period segmentation that depends on minimum set of parameters and requires no additional information on the expected shape of periods or nature of the data. Since biomedical signals may vary with respect to the person's physical abilities are generally characterised with irregularities in periodicity and time scales [13], the method of period extraction that is independent of signal shape or nature of the data would contribute to the fields of signal analysis and pattern recognition as well as their biomedical applications.

II. Related work

The basic methods of evaluating frequency of time series rely on approximating the time series with a sine model using least square estimation [14]. An asymptotically equivalent method is the maximization of the periodogram [15] of the time series. Though the estimates obtained through these methods converge to the true value of frequency for strictly periodic time series, these methods are generally not applicable for time series of non-stationary periodicity. However, one more often encounters the time series which, though not strictly periodic, demonstrate behavior similar to periodic. Such time series in various studies can be referred to as quasi-periodic or nearly periodic. We chose the term "nearly periodic" as less

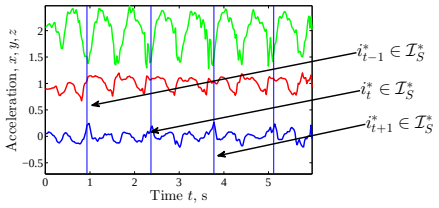
ambiguous. A formalization of this term will be given in the next Section.

Methods of estimating instantaneous frequency of a quasi-periodic signals are listed, for example, in [16] and include phase differencing, least square phase estimation, discrete-time Hilbert transform and others. For example, the authors of [17] estimate the fundamental (i.e., the lowest frequency) as a fixed point of a mapping based on Short Time Fourier Transform, minimizing the self-introduced Carrier-to-noise ratio. The paper [16] extends the existing methods for evaluating the instantaneous frequency of continuous signals for the case of discrete time series.

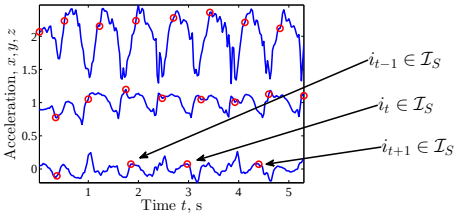
Another important class of period extraction methods stems from the ad hoc methods. Using specific information related to the nature and physical properties of the measured signal helps to enhance the results of period extraction. For example, the methods that aim to detect walking steps with accelerometer-based time series usually are designed to detect peaks in acceleration associated with heel strike or other phase of walking gait [18], [19]. The algorithms tested in the papers [18], [19] are derived from threshold methods when the step is count every time the amplitude of the signal exceeds a predefined or adaptive threshold [20]. The common methods also regarded in this papers include looking for a given pattern in a signal, window processing, transforming the signal into frequency domain [21], [22] and clustering the time- and frequency-domain features. More information on time- and frequency-domain features used in human motion analysis can be found in [23]. Though the paper [23] addresses the problem of classifying activities instead of segmenting one type of gait into cycle, the features that are used for solving this classification are good examples of application specific information about the processed data. These methods rely on a form of the signal and are based on detecting the peaks in acceleration. We aim to design a more general approach to the problem of extracting periods from the time series. To solve this problem, we exploit the singular value decomposition of time series. We construct the trajectory matrix of the studied time series, and compute its principal components. We describe a procedure of selecting a pair of principal components corresponding to the fundamental periodic. The ending points of the periods are defined through cutting the trajectory of the selected pair with its symmetry axis. In this paper the period extraction problem is closely connected to the interpretability of the segments. We define our purpose as partitioning nearly periodic biosignals into the segments that would be interpretable in a sense of nature of the signal.

III. Problem Statement

Let us explain the idea of nearly periodic time series on example of time series of human gait, such as walking, running or leaping. We suppose that each studied time series of human gait comprises a fundamental period, connected with the cycle of gait. During the cycle of normal human gait, each lower limb goes through a stance phase and a swing phase and then returns to the stance phase [24], [25]. We define the *fundamental period* of human gait as a segment of the time series, measured between two consequential moments of time when a person's body takes the same pose. The sequence of times series points, measured with accelerometer during one gait cycle is repeated as a human walks on. The time scale of the cycle and the exact values of the acceleration may change, but some characteristic shape of a cycle stays unchanged. For example, each cycle of time series of vertical component of acceleration, measured as human walks, contains two maximums: the first one corresponding to the landing of the heel and another one to pushing off the ground. The shape of period may be exposed to transformations from cycle to cycle, but its origin is the same. This property is what characterises nearly periodic time series as we define them.



(a) Ground-truth segmentation points $\mathcal{I}_S{}^*$



(b) Automatically extracted segmentation points $\mathcal{I}_S$

Fig. 1. The figure explains the notions of ground-truth segmentation points $i^*_t \in \mathcal{I}^*_S$ and the set $\mathcal{I}_S$ of extracted segmentation points i_t. In the subfigure (a) manually segmented time series are plotted; vertical lines divide the time range into periods. Blue line stands for a_x, red and green — for a_y and a_z. In the subfigure (b) the segmentation results are plotted. Red points mark the limits of the extracted segments. The figure demonstrates the inconvenience of using tri-axal accelerometer data: the periods extracted along different dimensions have to be compromised.

We formalize the idea of a nearly periodic time series as following. The nearly periodic time series $X = \{x(1), \dots, x(i), \dots, x(m)\}$ of length m is composed of a number of repetitions of the same pattern, exposed to slight shape changes and time scaling. Thus the nearly periodic time series X is defined by the set $\langle \mathbf{s}, a(i,s), f(i) \rangle$, where vector $\mathbf{s} = [s_1, \dots, s_T]^\mathsf{T}$ defines the basic shape of a period, the function $a(i,s), i \in \{1, \dots, m\}$ modifies this shape and the piecewise function $f(i) \mapsto \{1, \dots, T\}$ performs the time scaling. Then nearly periodic time series X can be expressed as

$$x(i) = a(i, s_{[f(i)]}). \tag{1}$$

Function f has a finite number of subdomains such that f is monotonically increasing from 1 to T on each subdomain. The points $\mathcal{I}^*_S = \{i^*_1, \dots, i^*_S\} = \{i^*_t \mid f(i_t + 1) < f(i_t)\}$ are considered the ending points of the periods. The set $\mathcal{I}^*_S$ is expected as the result of the period extraction procedure.

In this paper we solve the problem of period extraction of time series $X = \{x(i)\}_{i=1}^m$, regarding a result of pe-

riod extraction as a partition $\mathcal{I}_\mathcal{S} = \{i_1, \dots, i_S\}$ of time series X into sequence $\mathcal{S} = \{S_1, \dots, S_T\}$ of segments $S_t = [x(i_t), \dots, x(i_{t+1} - 1)]^\intercal$ corresponding to time-scaled and transformed basic shape. Fig. 1 presents an example of time series segmented into periods. The tree-dimensional acceleration time series of human walking gait are plotted: a_x in blue, a_y in red and a_z in green (Fig. 1(a)). This time series were measured by a smartphone accelerometer and then manually segmented into periods. The acceleration was measured at sampling rate of 40 points per second. The manually extracted segmentation points $i_t^* \in \mathcal{I}_S^*$ are called *ground-truth*. The ground-truth segmentation points are depicted on Fig. 1(a) with vertical blue lines. Fig. 1(b) presents the same time series, segmented automatically with the proposed method. The red points mark extracted segmentation points $i_t \in \mathcal{I}_S$. For our purposes it is unimportant where to start measuring the cycle of gait: when the heel was landing, or when it was taking off. It is only necessary that the cycle of walking starts and end with the same pose.

Throughout the paper we use the term *segmentation* referring to any procedure of partitioning time series into segments; the proposed method is referred to as *period extraction*. Similarly, the results of the segmentation in general are called *segments*; the results of period extraction are called *extracted periods*.

To evaluate the results of the period extraction procedure, we use the technique proposed in [26]. The authors of [26] propose to regard the results the segmentation problem as a two-class classification problem of the set of points $\{x(1), \dots, x(m)\}$. Once the ground-truth set $\mathcal{I}_\mathcal{S}^* = \{i_1^*, \dots, i_S^*\}$ of ending points of the segments is fixed, we define the classes for $x(i)$ as

$$x(i) \in \begin{cases} 1, \text{ if } i \in \mathcal{I}_\mathcal{S}, \\ 0 \text{ otherwise.} \end{cases}$$

Then the results of the segmentation can be assessed by calculating the accuracy of the classification problem, with some corrections. Depending on the sampling rate one might allow small deviations of the obtained ending points $i_t \in \mathcal{I}_\mathcal{S}$ from the ground-truth segmentation $\mathcal{I}_\mathcal{S}^*$. This means that if the segmentation point from $\mathcal{I}_\mathcal{S}$ is close enough to some ground-truth point i_t^* say, $|i_t - i_t^*| < \epsilon$, and no other points from $\mathcal{I}_\mathcal{S}$ are in the ϵ-neighbourhood of i_t^*, the ϵ-neighbourhood of i_t^* is said to consist of true negatives and one true positive. If $q \geq 1$ points from $\mathcal{I}_\mathcal{S}$ are found in $(i_t^* - \epsilon, \dots, i_t^*, \dots, i_t^* + \epsilon)$, then $q - 1$ of them are considered false positives. Alternatively, if $q = 0$, then $x(i_t^*)$ is counted as false negative. Since the sample for such problem obviously has more negatives than positives, the number of true negatives is not very informative and the adjusted coefficients are recommended in this case. In this paper we use F_1-score to estimate the quality of period extraction.

$$F_1 = \frac{2TP}{2TP + FN + FP}, \tag{2}$$

Here TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives respectively. The coefficient ranges from zero to one. The adequate segmentation results yield the values of F_1 close to one.

IV. Period extraction using principal components

In this section we describe the proposed method of period extraction.

A. Decomposing the time series with principal components analysis (PCA)

Let the time series X be a decomposition

$$X = \hat{X} + \tilde{X} + \boldsymbol{\varepsilon}, \tag{3}$$

where $\hat{X}$ is the trend, $\tilde{X}$ is the periodic time series or a sum of periodic time series, and $\boldsymbol{\varepsilon}$ is the noise. Each constituent of the decomposition (3) can be approximated with a high level of accuracy [27] with a combination of principal components of trajectory matrix $\mathbf{H}$

$$\mathbf{H} = \begin{pmatrix} x(1) & x(2) & \dots & x(N) \\ x(2) & x(3) & \dots & x(N+1) \\ \vdots & \vdots & \ddots & \vdots \\ x(m-N+1) & x(m-N+2) & \dots & x(m) \end{pmatrix}^\mathrm{T}$$

of the centered time series X

$$x(i) \mapsto x(i) - \frac{1}{m}\sum_{i=1}^{m} x_i, \quad i = 1, \dots, m$$

The parameter N, which determines the longest periodicity captured by PCA, is called *window length*. The method captures periodicity with period length T most accurately when N is multiple of T. Matrix $\mathbf{H}$ can be used to reconstruct the original time series X or its constituents trough anti-diagonal averaging. For example, to find an approximation of $\tilde{X}$ compute the singular value decomposition of covariance matrix of $\mathbf{H}$

$$\frac{1}{N}\mathbf{H}^\intercal\mathbf{H} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\intercal, \quad \Lambda = \mathrm{diag}(\lambda_1, \dots, \lambda_N) \tag{4}$$

and find principal components $\mathbf{y}_j = \mathbf{H}\mathbf{v}_j$, respective to positive eigenvalues of $\mathbf{H}^\intercal\mathbf{H}$. Each principal component $\mathbf{y}_j$ may be used to reconstruct a part of the trajectory matrix $\mathbf{H}$:

$$\mathbf{H} = \mathbf{H}_1 + \dots + \mathbf{H}_d, \quad \mathbf{H}_j = \sqrt{\lambda_j}\mathbf{v}_j\mathbf{y}_j^\intercal.$$

To approximate $\tilde{X}$, select several components $\mathbf{y}_j$ to compose matrix $\tilde{\mathbf{H}}$ and reconstruct the constituent $\tilde{X}$ of the original time series X via diagonal averaging of $\tilde{\mathbf{H}}$. This is the way to extract a periodic constituent. The problems is to select the principal components for $\tilde{\mathbf{H}}$. The following theorem states that a wide range of nearly periodic time series have two corresponding principal components with consequent numbers j and $j + 1$.

Theorem 1. *For time series*

$$X = \{x(1), \dots, x(i), \dots x(m)\}$$

of the form

$$x(i) = A_i \cos(2\pi w i + \phi) \tag{5}$$

with $w \in (0, 1/2)$, $\phi \in [0, 2\pi)$, $m \cdot w \in \mathbb{N}$ *and* $A_i : \exists C \in \mathbb{R} |A_i| < C \, \forall i$ *the principal components* $\mathbf{y}_1$ *and* $\mathbf{y}_2$ *can be expressed as*

$$y_1(l) = B_1(l)\cos(2\pi w l + \phi_1),$$

$$y_2(l) = B_2(l)\cos(2\pi w l + \phi_2),$$

$$\phi_1,\ \phi_2 \in [0, 2\pi),\ l = 1, \ldots, m - N + 1$$

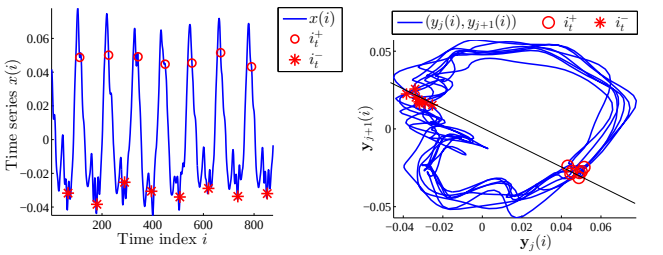
and the difference $|\phi_1 - \phi_2| \to \pi/2$.

The proof of this theorem can be found in the Appendix.

This result shows that each periodic constituent with $T > 2$ ($w < 1/2$) can be approximated with two principal components: the sine and cosine functions with the same frequency as the attached periodic. It is also shown in the Appendix that such principal components have asymptotically equal eigenvalues. This fact allows to consider only consequential pairs of principal components $(\mathbf{y}_j,\ \mathbf{y}_{j+1})$ in the procedure of automatical principal components selection, described in the Section "Automatical selection of a pair of principal components".

For the purpose of the period extraction we won't need to reconstruct the periodic constituent itself, it will be enough to select the pair of principal components $\mathbf{y}_j$ and $\mathbf{y}_{j+1}$ corresponding to the fundamental period and observe the trajectory $(\mathbf{y}_j,\ \mathbf{y}_{j+1})$.
We explain the procedure of selecting the optimal pair of principal components for period extraction in the Section "Automatic selection of the fundamental pair of principal components".

Note that the conditions of the Theorem hold for any nearly periodic time series with no time scaling. After time scaling the Eq. (5) only approximately describes the time series and the results of the Theorem are not necessarily valid. The impact of time scaling on the results of period extraction may vary.



(a) (b)

Fig. 2. An example of dissecting of a trajectory of a pair of principal components of time series X to obtain half-periods. Subfigure (a) presents the time series X and the extracted half-periods $S_t^- = [x(i_t^-), \ldots, x(i_{t+1}^+ - 1)]^\intercal$ and $S_t^+ = [x(i_t^+), \ldots, x(i_{t+1}^- - 1)]^\intercal$. Each separate set $\mathcal{I}_\mathcal{S}^-$ and $\mathcal{I}_\mathcal{S}^+$ defines a partition of time series into periods. Subfigure (b) presents the trajectory of principal components dissected by an arbitrary axis. Red circles and red stars denote clusters of points of approximately equal phase angle. These points form the segmentation sets $\mathcal{I}_\mathcal{S}^+$ and $\mathcal{I}_\mathcal{S}^-$.

B. Dissection of the phase trajectory with its symmetry axis

Consider some strictly periodic time series $X = \{x(i), i = 1, \ldots, m\}$ of period T:

$$x(i + T) = x(i).$$

According to the results, presented above $\tilde{X}$ has only two principal components $\mathbf{y}_1,\ \mathbf{y}_2$ with non-zero corresponding eigenvalues $\lambda_1 \neq 0,\ \lambda_2 \neq 0$: the sine and the cosine with the same period equal to T. The trajectory of normalized principal components $\mathbf{y}_1^*$ and $\mathbf{y}_2^*$

$$\mathbf{y}_j^* = \frac{\mathbf{y}_j}{\sqrt{\sum_{l=1}^{m-N+1} y_j^2(l)}},\ j = 1, 2$$

plotted against each other forms a spiral with fixed radius equal to 1 and center at (0,0):

$$y_1^*(l) \approx \cos(2\pi w l + \phi),$$

$$y_2^*(l) \approx \sin(2\pi w l + \phi).$$

We will further suppose that principal components are normalized.

One loop of the trajectory corresponds to a complete period of the each principal component or, equivalently, a complete period of initial time series, since their periods are equal. Thus the points of time series $\tilde{X}$ with the same coordinates in (y_1, y_2) space correspond to the same phase angle.

To find the ending points of the periods we cut the trajectory of a pair of principal components and cut it with a line that crosses the coordinate center ($y_2 = ky_1$ or $y_1 = 0$). Cutting the trajectory along this axis we obtain a partition $\mathcal{I}_\mathcal{S} = \{i_1^-, i_1^+, \ldots, i_T^-, i_T^+\}$ of the time series into negative $S_t^- = [x(i_t^-), \ldots, x(i_{t+1}^- - 1)]^\intercal$:

$$y_2(i) - ky_1(i) < 0 \quad \text{for all } i \in \{i_t^-, \ldots, i_{t+1}^- - 1\}$$

$$(y_1(i) < 0 \quad \text{for all } i \in \{i_t^-, \ldots, i_{t+1}^- - 1\}), t \leq T/2$$

and positive half-periods $S_t^+ = [x(i_t^+), \ldots, x(i_{t+1}^+ - 1)]^\intercal$:

$$y_2(i) - ky_1(i) > 0 \quad \text{for all } i \in \{i_t^+, \ldots, i_{t+1}^+ - 1\}$$

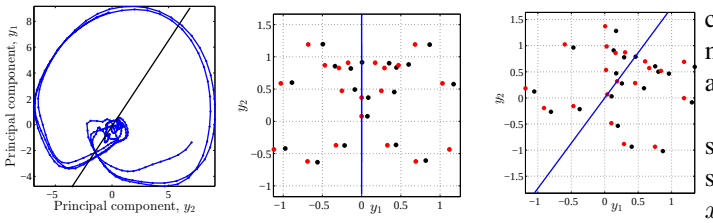
$$(y_1(i) > 0 \quad \text{for all } i \in \{i_t^+, \ldots, i_{t+1}^+ - 1\}), t \leq T/2.$$

Assume $i_t^- < i_t^+$, than $i_t^- = i_t^+ - 1$. Joining half-periods S_t^- and S_t^+, we obtain segmentation $\mathcal{I}_\mathcal{S} = \{i_1^-, \ldots, i_T^-\}$ into fundamental periods. Thus we obtain half-periods that will be later merged into periods.

The Fig. 2 demonstrates an example of implementing this procedure. Fig. 2 (a) represents nearly periodic time series X defined by a set $\langle \mathbf{s}, a, f \rangle$. The blue line correspond to the processed time series. Red circles and stars correspond to the extracted ends of half-periods: i_t^+ and i_y^-, respectively. A pair $(\mathbf{y}_j, \mathbf{y}_{j+1})$ of optimally chosen (for now let us omit the question of selecting a pair of components) principal components for X is fixed. The trajectory of selected components is plotted in Fig. 2, (b). Due to the application of time scaling f and shape transformation a, lengths of periods vary over the time. This change results in the trajectory of the normalized components forming a two-dimensional spiral with slightly varying radius (rather than a constant radius) with various

number of time points per loop (instead of fixed number of points equal to period length). Now the points from different loops that correspond to the same phase angle form clusters in $(y_j,\ y_{j+1})$ space, instead of having the same coordinates as it was the case with strictly periodic time series. When we cut the trajectory, plotted in blue, with a black axis the trajectory splits into "above" and "below" parts, each formed by half-loops, which correspond to half periods. Sets $\mathcal{I}_{\mathcal{S}}^{+}$ and $\mathcal{I}_{\mathcal{S}}^{-}$ of ending points of this half-loops are the candidates for $\mathcal{I}_{\mathcal{S}}$, each leading to a distinct partition of time series into segments. Fig. 2 (a) presents both options of period extraction with ending points marked as red circles for $\mathcal{I}_{\mathcal{S}}^{+}$ and red stars for $\mathcal{I}_{\mathcal{S}}^{-}$. The difference between these partitions lies in the variance of phase angle for sets $\{x(i_t^{+})\}$ and $\{x(i_t^{-})\}$: since the time series are not strictly periodic, the points $x(i_t),\ i_t \in \mathcal{I}_{\mathcal{S}}$ do not have the same coordinates $(y_j(i_t), y_{j+1}(i_t))$ and correspond to different phase angles. In order to minimize this variance, we choose the set $\mathcal{I}_{\mathcal{S}}^{\bullet}$ that provides the smaller distance between the respective points $(y_j(i_t^{\bullet}), y_{j+1}^{\bullet})$:

$$\mathcal{I}_{\mathcal{S}} = \underset{\bullet\in\{+,-\}}{\operatorname{argmin}} \sum_{i_t^{\bullet}, i_q^{\bullet}\in\mathcal{I}_{\mathcal{S}}^{\bullet}} ||\big(y_j(i_t^{\bullet}), y_{j+1}(i_t^{\bullet})\big) - \big(y_j(i_q^{\bullet}), y_{j+1}(i_q^{\bullet})\big)||.$$



(a) (b) The axis is known (c) The axis is unknown

Fig. 3. (a) An example of a trajectory of a pair of principal components, dissected with its symmetry axis. (b) Symmetrization of a set of points (black) with respect to a known axis covering the coordinate center (vertical line). The resulting symmetrized set is plotted in red. (c) The results of the same set with unknown axis. The position of axis is optimized to minimize the distance from the original set to the symmetrized.

It can be inferred from Fig. 2, the results of period extraction depend highly on the slope of the cutting line. We introduce the following heuristic for the axis slope selection: let us split the trajectory of the principal components with the axis that is the closest to the symmetry axis of the dissected trajectory as shown in Fig. 3(a). We do so wishing for sets $\mathcal{I}_{\mathcal{S}}^{+}$ and $\mathcal{I}_{\mathcal{S}}^{-}$ to differ as much as possible in respect to the phase variance in order to have the ability to choose. To find the symmetry axis we exploit the method of symmetrisation of a set of points proposed in [28], [29]. Let the symmetry axis coincides with the ordinates axis $y_1 = 0$. According to [28], [29], symmetrize a set of points $(\mathbf{y}_1, \mathbf{y}_2)$ an auxiliary vector $Y = [\mathbf{y}_1^{\mathrm{T}}, \mathbf{y}_2^{\mathrm{T}}]^{\mathrm{T}}$ is formed. Let the elements of Y be ordered so that for a symmetrized vector Y_s it holds:

$$y_1(i) = -y_1(m'+i),\ i = 1, \ldots, m',$$

$$y_2(i) = y_2(m'+i),\ i = 1, \ldots, m',$$

$$y_1(i) = 0,\ i = 2m'+1, \ldots, m.$$

Then the symmetrization that minimizes the deviation

$$||Y_s - Y||_2$$

of symmetrised vector Y_s from the original vector Y is obtained through the transformation:

$$Y_s = \mathbf{Q}Y,$$

where

$$Q = \frac{1}{2}\mathbf{I}_{2n} + \frac{1}{2}\begin{pmatrix} -S & 0 \\ 0 & \mathbf{S} \end{pmatrix}, \quad \mathbf{S} = \begin{pmatrix} 0 & \mathbf{I}_m & 0 \\ \mathbf{I}_m & 0 & 0 \\ 0 & 0 & \mathbf{I}_{n-2m} \end{pmatrix}.$$

In general the symmetry axis is given by $y = \mathrm{tg}(\phi)x$ or $x = 0$. We'll define the angle ϕ as the solution of a minimization problem

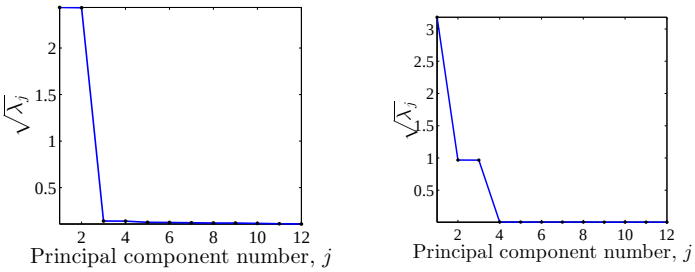
$$\phi = \underset{\phi\in[0,\pi/2]}{\operatorname{argmin}} ||(Y_\phi)_s - Y_\phi||,$$

where Y_ϕ is the vector of coordinates of Y in the rotated by ϕ coordinate system. Fig. 3(b) presents a set of points and the results of its symmetrization with respect to the ordinates axis. The original positions of the points are marked with black dots. The symmetry axis is plotted in blue. The red dots correspond to the adjusted positions of the points: the set of red points in symmetrical with respect to the ordinates axis.

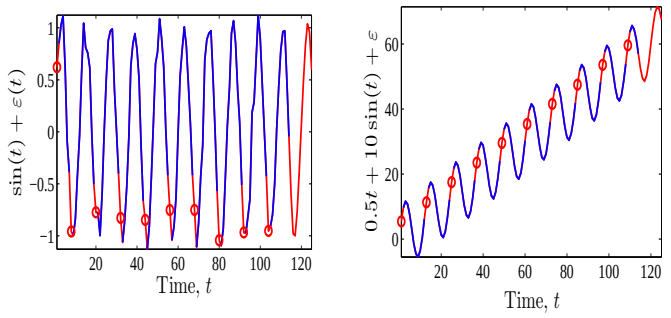
The original set of (black) points was rotated around the coordinates centre and then symmetrized. The results of symmetrization of the rotated set and the evaluated symmetry axis are presented by the Fig. 3(c).

Example 1: period extraction on synthetic data. To demonstrate the proposed method of period extraction, let us consider two synthetic time series: $x_1(t) = \sin(i) + \varepsilon$ and $x_2(i) = 0.5i + 10\sin(i) + \varepsilon$. To extract the periods we depict each time series X_1 and X_2 with a pair of selected principal components and cut the trajectory of $(\mathbf{y}_j, \mathbf{y}_{j+})$ with its symmetry axis to determine segmentation points $\mathcal{I}_{\mathcal{S}}$. The simplest way to select a pair of principal components is visual analysis: for a pair with approximately equal eigenvalues we plot the trajectory of principal components. If the trajectory is similar to the spiral, then we select it as a periodic pair. The results of period extraction of the time series $x(i) = \sin(i) + \varepsilon$ and $x_2(i) = 0.5i + 10\sin(i) + \varepsilon$ based on the pair of principal components selected this way are presented on the Fig. 4. The time series $x_1(i)$ are plotted in Fig. 4(c) and Fig. 4(d). Fig. 4(a) and (b) demonstrate how the values $\sqrt{\lambda_j}$ decrease with j. The time series $x_1(i)$ only contain a periodic constituent. The first two components were selected since they have almost equal eigenvalues that are significantly greater than zero. The time series $x_2(i) = 0.5i + 10\sin(i) + \varepsilon(i)$ consists of the periodic $\tilde{x}_2(i) = 10\sin(i)$ and of the trend $\hat{x}_2(i) = 0.5i$ (see Fig. 4(d)). In this case the first principal component $\mathbf{y}_1$ correspond to the trend $\hat{X}_2$, while the second and the third have approximately equal eigenvalues and correspond to the periodic $\tilde{X}_2$. The red circles in Fig. 4(c) and Fig. 4(d) correspond to the extracted segmentation points i_t.

Example 2: principal component selection for the fundamental period extraction on the data from [12]. Although the visual analysis works well for this simple example, this method, but from other evident faults, has a limitation

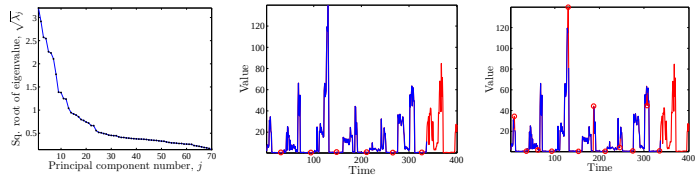


(a) (b)



(c) $x_1(i) = \sin(i) + \varepsilon$ (d) $x_2(i) = 10\sin(i) + 0.5i + \varepsilon$

Fig. 4. An example of period extraction for the synthetic time series $x_1(i) = \sin(i) + \varepsilon$ and $x_2(i) = 0.5 + 10\sin(i) + \varepsilon$ on the basis of the first pair of principle components with approximately equal eigenvalues. The squared values of eigenvalues are given in subfigures (a) and (b). The pairs $(\mathbf{y}_1, \mathbf{y}_2)$ and $(\mathbf{y}_2, \mathbf{y}_3)$ were chosen for (a, c) and for (b, d) respectively. The subfigures (c) and (d) demonstrate the results of period extraction: red circles mark the segmentation points (ends extracted of periods).



(a) (b) (c)

Fig. 5. The figure demonstrates a limitation of visual analysis of principal components. The subfigures plot: (a) the eigenvalues of trajectory matrix and (b, c) the results of period extraction based on the first (b) and the second (c) periodic pairs $(\mathbf{y}_1, \mathbf{y}_2)$ and $(\mathbf{y}_3, \mathbf{y}_4)$ of principal components. The subfigure (a) suggests several candidate-pairs with approximately equal eigenvalues, each pair leading to a different segmentation result.

connected with the need to chose between several periodic components. One possible solution is to chose the pair with maximum eigenvalues, assuming that principal components with lower eigenvalues approximate periodic components with higher frequencies. This approach was implemented to the data [12], consisting of several periodic constituents. The results of processing and segmenting the time series are pictured on Fig. 5. Fig. 5(a) and (b) plots the values of the square roots $\sqrt{\lambda_j}$ of the eigenvalues. It can be seen that each time series has several pair-candidates: there is more than one pair of principal components $\mathbf{y}_j, \mathbf{y}_{j+1}$ with approximately equal eigenvalues $\lambda_j \approx \lambda_{j+1}$. Since each pair correspond to different periodicity, the results of segmentation into periods depend on a selected pair. Fig. 5(b) and Fig. 5 (c) demonstrate the results of segmentation of the time series based on the first $(\mathbf{y}_1, \mathbf{y}_2)$ (b) and the second $(\mathbf{y}_3, \mathbf{y}_4)$ (c) periodic pairs for each time series. The blue line is the historical time series, the red circles mark the starting/ending points of the segments. The red line correspond to a part of the time series that did not fit into the trajectory matrix $\mathbf{H}$. This experiment demonstrates that the time series might have several candidate pairs even if the periodicity can not be seen by eye. The principal components must be chosen upon some formal criterion.

V. Automatic selection of the fundamental pair of principal components

The procedure of automatic selection utilized in the paper is based on the method proposed in the paper [30]. The idea behind this method is to detect principal components, periodic with the same frequency comparing their spectral densities.

For time series X the Digital Fourier Transform is given by:

$$x(i) = \sum_{k=1}^{m} f(k) \mathrm{w}_m^{-(k-1)(i-1)}, \quad \mathrm{w}_m = \exp(-2\pi \mathrm{i}/m). \quad (6)$$

In this paper the spectral densities are computed with Fast Fourier Transform. Fig. 6(a, b, c) pictures the spectral density of the time series X: the dependencies of the amplitudes $|f(k)|$ on the frequencies $(k-1)/m$, where

$$f(k) = \frac{1}{m} \sum_{i=1}^{m} x(i) \mathrm{w}_m^{-(k-1)(i-1)}.$$

Knowing the spectral density gives us information about the periodicities that are most strongly manifested in time series X. We choose a minimum nonzero frequency $w_{\min}$, presented in the time series as a basic approximation of the fundamental frequency to look for when choosing a periodic pair of principal components.

The procedure of selecting a pair of principal components is the following.

1) Form a set $\mathcal{Y}$ of candidate pairs. Since the eigenvalues of the principal components of periodic pair are almost equal, the principal components in candidate pairs have consequent indices:
$$\mathcal{Y} = \{(\mathbf{y}_j, \mathbf{y}_{j+1}) | \lambda_j + 1 > 0,\ 1 \le j \le N.\}$$
2) As it was shown in Section "Principal component selection" a periodic pair of principal components consists of two periodic functions differing only in their phase. The spectral densities of such principal components have their maximums of amplitude $|f(k)|$ in the same frequencies $(k-1)/m$. For each candidate pair $(\mathbf{y}_j, \mathbf{y}_{j+1}) \in \mathcal{Y}$ we check if the maximums of their spectral densities coincide to decide whether this candidate pair indeed forms a periodic pair.
Let k_j be the number of frequency, corresponding to the maximum value $|f^j(k)|$ of principal component $\mathbf{y}_j$
$$k_j = \operatorname*{argmax}_{k \in \{0,\dots,m-1\}} |f^j(k)|.$$

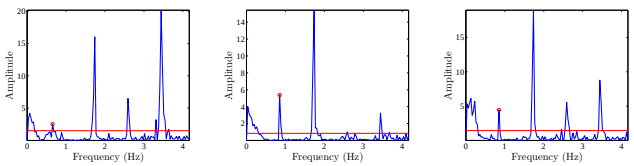
Due to the noisiness or signal modulations, the spectral density may become "smeared": the peaks become less sharp and precise and the arguments k_j and k_{j+1} may not exactly coincide even for matching principal component. We construct the set $\mathcal{Y}^*$ of periodic candidate pairs $(\mathbf{y}_j, \mathbf{y}_{j+1}) \in \mathcal{Y}$, verifying that

$$|k_j - k_{j+1}| \leq \varepsilon M,$$

where M equals to the number of samples per second. Parameter ε controls accepted smearing of spectral density. The results of computational studies, provided in [30] allow us to fixate this parameter equal to $1/M$. We then obtain a set of periodic candidate pairs:

$$\mathcal{Y}^* = \{(\mathbf{y}_j, \mathbf{y}_{j+1}) \in \mathcal{Y} | \; |k_j - k_{j+1}| \leq 1\}.$$

3) Among the candidate pairs obtained by the algorithm we choose a pair $(\mathbf{y}_j, \mathbf{y}_{j+1})$ with $(k-1)/m$ closest to minimum nonzero frequency $w_{\min}$ of time series X.



(a) (b) (c)

Fig. 6. Spectral density of the studied time series a_x, a_y, a_z (acceleration along the dimensions x, y, z). For each time series the first nonzero frequency peak $w_{\min}$ is detected. Then the pair of principal components with the closest frequency is chosen.

VI. Experimental study of PCA period extraction

In this section we describe several tests of performance of the proposed method in dependence on the shape $\mathbf{s}$ and the functions f and a of the nearly periodical time series X. All the experiments were conducted on synthetic data. In the experiments it is supposed that though the frequency and the shape of the signal may undergo spontaneous changes, these changes are not too drastic, so that the time series can be represented as distorted periodical time series. The exact range of "acceptable" change is hard to define due to the low level of formalization of the notion of nearly periodic time series we use. Here we investigate limits of "acceptable" changes experimentally. We use noisiness to model both shape distortions and time scaling and use the variance of noise as a measure of distortion.

A. Robustness to noise level

Nearly periodical time series X, used in the experiments, were obtained from strictly periodical time series $\tilde{X}$, $\tilde{x}(i+T) = s_i$, $1 \leq i \leq T$ with period T and shape $\mathbf{s}$ by exposing $\tilde{X}$ to different types of noise:

1) Additive noise

$$x(i+T) = \tilde{x}(i) + \varepsilon \tag{7}$$

with $\varepsilon \sim \mathcal{N}(0, \sigma)$ and noise level measured as $\sigma_\varepsilon / \max \tilde{X}$.

2) Argument noising

$$x(i+T) = \tilde{x}(i+\varepsilon), \tag{8}$$

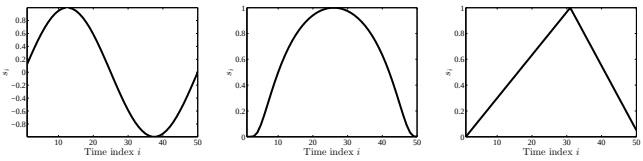
with $\varepsilon \sim \mathcal{N}(0, \sigma)$ and noise level measured as σ_ε / T.

3) Mixture-type noising

$$x(i+T) = x_2(i) \cdot (1+\varepsilon), \tag{9}$$

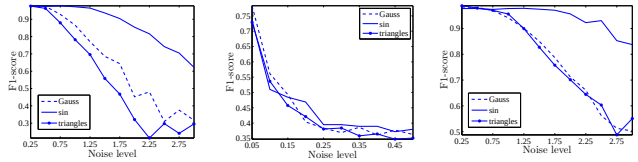
where X_2 is the result of applying type 2 (8) noise to the time series X. Here $\varepsilon \sim \mathcal{N}(0, \sigma)$ and noise level is measured as $\sigma_\varepsilon / \max \tilde{X}$.

Fig. 8 contains the dependencies of the F_1-score on the noise level, plotted for different shapes $\mathbf{s}$, for type 1 (7), type 2 (8) and type 3 (9) of exposure to noise. The basic shapes used in this experiment are the sine, the Gaussian bell and the triangle, plotted on Fig. 7 (a, b, c). Each shape $\mathbf{s}$ had length $T = 50$ and was repeated $S = 30$ times.



(a) sine (b) Gauss (c) triangle

Fig. 7. The examples of basic shapes used in experiments to generate nearly periodical time series.



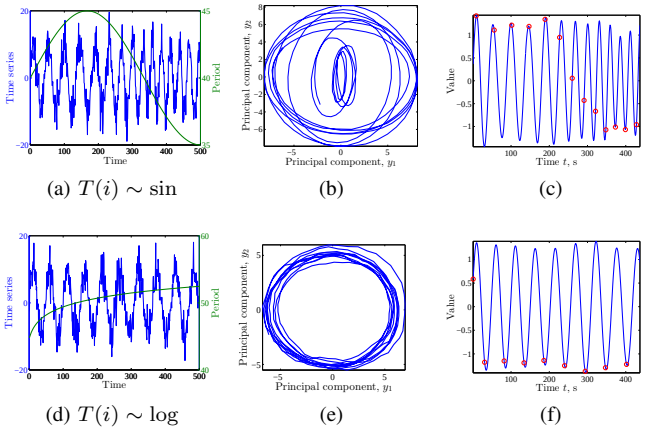
(a) Type 1 (b) Type 2 (c) Type 3

Fig. 8. The dependencies of F_1-score on the noise level for three types of noising. Each subfigure corresponds to a type of noising and contains the dependencies for three basic shapes: sine, Gaussian bell and a triangle.

The experiments provide some insight into the range of acceptable transformations: performance of the proposed method is adequate ($F_1 \gtrsim 0.7$) when $\sigma_\varepsilon \lesssim 1.5 \max(X)$ for types 1 and 3 of noising, and $\sigma_\varepsilon \lesssim 2T$ for type 2.

B. Increasing Robustness to frequency modulations with moving window technique

Fig. 9 demonstrates how different phase modulations affect the results of period extraction. Fig. 9(a) and 9(d) plot the time series (in blue) with sine-modified and log-modified period $T(i)$ respectively. The modulation functions are plotted on the same figures in green. The changes of period are more rapid for the time series in Fig. 9(a) than for the time series in Fig. 9(d). The trajectory (Fig. 9(b)) of periodic principal components of the first time series is less stable compared

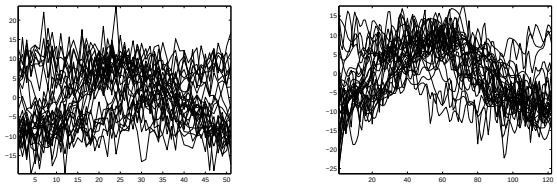


(a) $T(i) \sim \sin$ (b) (c)

(d) $T(i) \sim \log$ (e) (f)

Fig. 9. The figure illustrates two examples of segmenting phase-modulated time series. The phase is given by $2\pi i/T(i)$ with $T(i) \sim \sin(2\pi i/m)$ for subfigures (a, b, c) and $T(i) \sim \log(1+i)$ for (d, e, f). In the second case the change of period length is smooth and the method extracts the segmentation points of approximately the same phase. In the first case, the changes are much more significant. As a result, the trajectory of principal components changes its shape and we find considerable phase shift in segmentation points. Splitting the trajectory into inner and outer circles will improve the results, which is the motivation for applying moving window technique.

to the trajectory of the second time series (Fig. 9(e)). As a result, one can see the continuous shift in the phase of the segmentation points $x(i_t)$, $\imath_t \in \mathcal{I}_{\mathcal{S}}$ for the first time series (Fig. 9(c), blue line plots the time series, red circles mark the segmentation points $x(i_t)$) and almost no shift for the second time series (Fig. 9(f)). To reduce this effect we apply the moving window technique. The details of the technique can be found in [31]. The point of the method is to split the time series into shorter parts, so that the each part would have more stable periodicity and the trajectory of principal components would have more regular structure. The results of applying moving window technique to period extraction from the time series with sine-modified period from Fig. 9(a) are demonstrated by Fig. 10(a) and Fig. 10(b), which plot the extracted periods, aligned to the same length for the basic version of the proposed method and its modification with moving window technique. Fig. 10 shows less diversity in the second case. The value of F_1 also improved.



(a) "before": $F_1 = 0.4997$ (b) "after": $F_1 = 0.6414$

Fig. 10. The figure displays the results of period extraction for a frequency-modulated time series with $T(i) \sim \sin(2\pi i/m)$. Subfigure (a) presents the extracted periods S_t, obtained with the basic version of the proposed method. The segments S_t were aligned to the same length. The continuous change of period length evoked the phase shift which can be seen as variety of form of the extracted periods. Subfigure (b) presents the aligned periods extracted with the moving window modification of the proposed method. The segments are now more alike due to reduced. For quantitative assessment of improvement, the values of F_1 for $\epsilon = 0$ are given.

In the experiments, the moving window technique was implemented if the length of the time series exceeded 10 periods lengths. We used $T_0 = 1/w_{min}$ as a zero-order approximation of period length T. The required length m of time series is at least $3T_0$. The complete procedure of period extraction sums up as following:

1) For the shorter time series $X = \{x(1), \dots, x(10T_0)\}$ find a new T_0 approximation of T and compute the list $\{\mathbf{y}_1, \dots, \mathbf{y}_N\}$ of principal components with $\lambda_1 \geq \dots \geq \lambda_N$. Here we set parameter N of the trajectory matrix $\mathbf{H}$ to $1.75T$, because the computational experiments demonstrated almost no dependence of the quality of period extraction on this parameter for $N \geq T$ (where T is the average period length). Choosing a small value of N one risks to miss the periodicities of higher frequency. Setting N too high increases computational time required to select a pair of principal components.
2) Chose a fundamental periodic pair $(\mathbf{y}_j, \mathbf{y}_{j+1})$.
3) Plot the trajectory of $(\mathbf{y}_j, \mathbf{y}_{j+1})$, cut it with the symmetry axis to obtain the segmentation points $\mathcal{I}_{\mathcal{S}}$.
4) Compute a new approximation of period length, $T_{PC} = m/(k_j - 1)$. This approximation is used to shift the sliding window.

If the time length if the time series is more than $10T_0$, the window is shifted to T_{PC}, so that the procedure is repeated for the time series $X = \{x(T_{PC}), \dots, x(T_{PC} + 10T_{PC})\}$, then the sliding window is shifted again and so on. After each shift the window captures a new piece of the original time series, extracting new segmentation points and computing new approximation T_{PC} to use on the next step. The procedure repeats until all m points of time series X are covered by the method[1].

VII. Analysis

In this section we describe some classical methods of estimating the period length and provide a comparison of the results we obtained applying the proposed method and the classical alternatives to the walking data set.

A. Alternative methods of estimating the period length

The Least Squares Estimation. To obtain the least squares estimation one fits the model

$$x(i) \approx \sum_{q=1}^{Q} A_q \cos(2\pi w_q i) + B_q \sin(2\pi w_q i)$$

best according to the residual sum of squares

$$\{\mathbf{w}_{\text{LSE}}, \mathbf{A}, \mathbf{B}\} = \underset{\mathbf{w},\mathbf{A},\mathbf{B}}{\operatorname{argmin}} \sum_{i=1}^{m} \left(x(i) - \sum_{q=1}^{Q} \big(A_q \cos(2\pi w_q i) + B_q \sin(2\pi w_q i)\big) \right)^2.$$

[1]The code for conducting experiments is available at http://svn.code.sf.net/p/mlalgorithms/code/ in the directory Group874/Motrenko2014TSsegmnetation/web/. There is also a demonstration version, available at http://193.233.212.81/Motrenko2014PeriodExtraction/start.html

Consider for simplicity the case of $q = 1$. Then $T_{\text{LSE}} = M/2\pi w_{\text{LSE}}$ and, as shown in [14], the residual squares functional can be modified to an asymptotically equivalent one (for $0 < w < 0.5$)

$$\{w_{\text{LSE}},\, A,\, B\} = \underset{w,\, A,\, B}{\operatorname{argmin}} \left\{ \sum_{i=1}^{m} x^2(i) - 2 \sum_{i=1}^{m} x(i) \big(A \cos(2\pi w i) + B \sin(2\pi w i) \big) + \frac{2}{m}(A^2 + B^2) \right\} \tag{10}$$

to provide the following estimates $\hat{A}$ and $\hat{B}$ of A and B

$$\hat{A}(w) = \frac{2}{m} \sum_{i=1}^{m} x(i) \cos(2\pi w i), \quad \hat{B}(w) = \frac{2}{m} \sum_{i=1}^{m} x \sin(2\pi w i).$$

Maximization of the periodogram. Note that the functional (10) can be written as following

$$\sum_{i=1}^{m} x^2(i) - 2 \sum_{i=1}^{m} \big(A \cos(2\pi w i) + B \sin(2\pi w i) \big) + \frac{2}{m}(A^2 + B^2) = = \sum_{i=1}^{m} x^2(i) - \frac{2}{m} \left| x(i) e^{-\mathbf{i} 2\pi w i} \right|,$$

which means the minimization (10) is equivalent to maximizing the periodogram [15] $\frac{2}{m} \left| x(i) e^{-\mathbf{i} w i} \right|$. In this paper we do not distinguish between these estimators, considering only the one given by (10) and treating it as least squares estimation. *The cross correlation estimation.* To obtain the cross correlation estimate, we partition the time series into S segments $S_t = \{x(i_{T(t-1)+1}, \ldots, x(i_{T \cdot t}))\}$ of same length T in the predefined range $[T_{\min}, T_{\max}]$ and compute the average Pearson's correlation coefficient

$$\text{Corr}(X, T) = \frac{1}{S-1} \sum_{t=1}^{S-1} \rho(S_t, S_{t+1})$$

between the neighboring segments S_t, S_{t+1}. The cross correlation estimate of the period is then given by

$$T_{Corr} = \underset{T}{\operatorname{argmax}}\, \text{Corr}(X, T).$$

We chose the $|T_{\max}| = m/4$, meaning that the time series should contain at least four periods, and the minimum number $T_{\min}$ of points in a period to three.

B. Comparing the results

The proposed method was tested on synthetic and real data. The evaluation technique exploited in this paper requires the knowledge of ground-truth segmentation. To apply it one needs the time series that are already marked. To avoid confusion, we do not use existing automatical step-detectors to obtain ground-truth segmentation data and only use manually segmented time series for comparison. This complicates the process of collecting data. This is why we had to use the synthetic time series for evaluation to increase the sample size.

According to our definition (1), each synthetic nearly periodical time series is defined by its shape $\mathbf{s}$, time scaling function f and shape-modifying function a. The basic shapes used in the experiment are sine (fig. 7(a)), Gaussian bell (fig. 7(b)) and triangle (fig. 7(c)). The f and a modifications used in the data generation are limited to the three types of noising specified by equations (7), (8), and (9). Synthetic time series were generated with random noise component ε sampled from $\mathcal{N}(0, \max \tilde{X})$ for type 1 and type 3 or $\mathcal{N}(0, T)$ for type 2.

The dataset includes accelerometer-based time series of human gait we measured with a hand-held smartphone. We used the accelerometer with sampling rate of 40 samples per second. The tri-axial data was merged into one-dimensional time series:

$$x(i) = \sqrt{a_x^2(i) + a_y^2(i) + a_z^2(i)}, \quad i = 1, \ldots, m.$$

We also used ECG time series and blood pressure time series from PhysioNet database [32]. The descriptions of the data sets cam be found in [33] for ECG time series and in [34] for the blood pressure. Each record was splitted into several one-minute long time series.

TABLE I
THE STRUCTURE OF THE SAMPLE SET.

Synthetic time series: total $N_{ts} = 240$				
	N_{ts}	m	S	$100\sigma_T/T$
Type 1	80	1500	30	0
Type 2	80	1500	30	7.2
Type 3	80	1500	30	6.8
Real-life time series: total $N_{ts} = 560$				
	N_{ts}	m	S	$100\sigma_T/T$
Walking	10	3000	18	5.7
Jogging	12	3000	20	5.2
Leaping	8	3000	15	6.9
ECG	350	6000	70	5.6
Blood pressure	180	7500	60	4.2

General statistics of each data set (synthetic time series, gait, ECG and blood pressure) that we used to run computational experiments is given in Table I. General statistics for synthetic data set are presented by types of noisiness. Gait data set is divided by particular gait instances: walking, jogging and leaping. Table I contains number N_{ts} of time series in each data set, average length m of time series and number S of periods per time series of this data set.

The rightmost column contains a random variability coefficient $100\sigma_T/T$ of period length T, expressed in percents. Here T is the average period length of time series of a particular data set, and σ_T is an estimate of change rate of period length

$$\sigma_T = \frac{1}{S} \sum_{t=1}^{S} \big| (i_{t+1}^* - i_t^*) - (i_t^* - i_{t-1}^*) \big|,$$

averaged by all time series for this data set. When the period length T does not change, $i_t^* \in \mathcal{I}_S^*$ are linear by t and $\sigma_T = 0$. If period length changes randomly with mean value T, than σ_T is an estimate of its standard deviation. We choose σ_T over standard deviation to measure random variability since it

compares only contiguous periods and is less affected by deterministic trend, if present. Variability coefficient $100\sigma_T/T$ for synthetic time series depends on the type of noisiness. Among the real life data sets the values $100\sigma_T/T$ are approximately the same with the lowest value for blood pressure data set and the highest value for the gait data.

Table II compares the proposed method with the alternatives presented in Section VII-A. For each data set we provide the results of period extraction with the proposed method, labeled "PCA", and the alternatives. The alternative methods are labeled "LSE" for Least Square Estimation or and "Corr" for the correlation method. As we have mentioned in Subsection VII-A, periodogram maximization is asymptotically equivalent to the least square estimation. Comparing the methods, we observed very similar performance for these methods. That is why "LSE" columns now present results for both of these methods, containing the best outcome of two options. Table II presents F_1 scores,

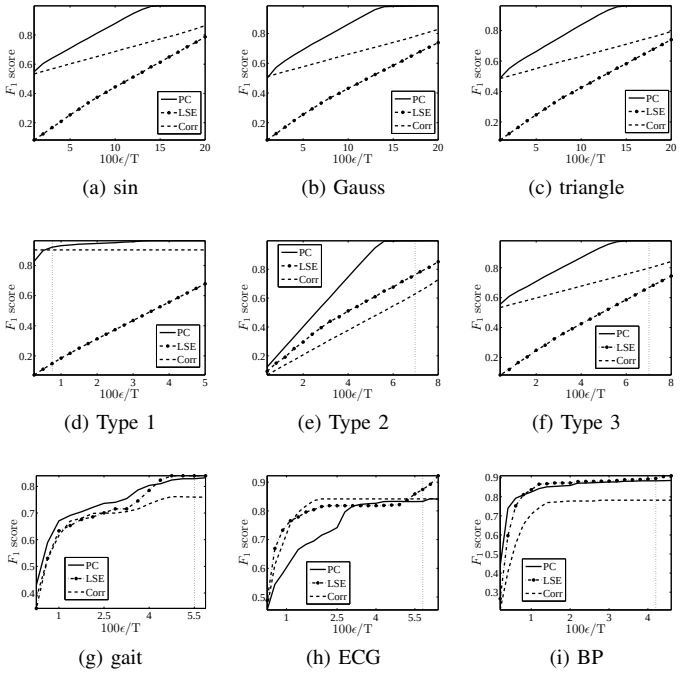
$$\text{Precision} = \frac{TP}{TP+FP}, \quad \text{and Recall} = \frac{TP}{TP+FN}$$

for the tested methods. Better performance yields higher values of these three criteria. The values presented in Table II were obtained by averaging F_1 scores, Precision and Recall over all time series from the particular data set.

The quality estimates F_1 depend on the value of uncertainty parameter ϵ, which defines how far the extracted segmentation points i_t can deviate from the "true" segmentation points i_t^* and be still considered correctly extracted. Fig. 11 illustrates how the value of F_1 score grows with the increase of ϵ. Fig. 11(a, b, c) and Fig. 11 (d, e, f) show $F_1(\epsilon)$ dependencies for the three tested methods applied to synthetic time series of different basic shapes and types of noisiness, accordingly. Fig. 11(g, h, i) correspond to gait, ECG and blood pressure time series. Horizontal axis represents uncertainty parameter ϵ, scaled by $100/T$ for easier comparison with $100\sigma_T/T$ values, presented in Table I.

As the value of ϵ increases, more segmentation points i_t are recognized as true positives. The number of false negatives does not change. The number of false negatives may only decrease by the same number as true positives have increased, resulting in constant sum $TP + FN + FP$. Best performing algorithms are the ones with higher areas under curve. Rapid convergence indicates low values of $FN+FP$. Almost linear behavior suggests that $TP+FN+FP$ is rather high compared to TP, which is not desirable. The Fig. 11 also demonstrates that performance of the tested methods depend much more on the type of noisiness than on the shape of a signal: the curves on Fig. 11(a, b, c) appear the same, while Fig. 11(d, e, f) differ considerably. That is why Table II presents the results by types of noisiness, rather than by shapes.

The exact value of uncertainty parameter ϵ was specified based on the sampling rate and the noisiness of the data, so that the ϵ would approximately equal to σ_T. Under this heuristic rule, an algorithm that extracts periods of constant length should yield a relatively high score even if there are considerable random variations in ground truth period lengths, but would fail to obtain perfect score if period length changes



(a) sin (b) Gauss (c) triangle

(d) Type 1 (e) Type 2 (f) Type 3

(g) gait (h) ECG (i) BP

Fig. 11. The dependence of F_1-score of the tested methods on the of the size of ϵ-neighbourhood of the ground truth segmentation points i_t^* for each type of synthetic time series. The size is given in percents of the average period length T. As the parameter ϵ increases, more extracted points i_t are recognised as true positives and the value of F_1 increases as well. Vertical lines mark the values of $100\epsilon/T$ that yield F_1, Precision and Recall values, reported in Table II.

TABLE II
AVERAGED PERFORMANCE OF THE TESTED METHODS: F_1-SCORE, PRECISION AND RECALL.

	F_1			Precision			Recall		
	PC	LSE	Corr	PC	LSE	Corr	PC	LSE	Corr
Type1	**0.9220**	0.1485	0.9039	**0.9623**	0.7861	0.6120	**0.9561**	0.7125	0.8966
Type2	**0.9953**	0.7825	0.6495	**0.9022**	0.7306	0.6070	0.9012	0.7094	**0.9930**
Type3	**0.9803**	0.6508	0.7891	**0.9466**	0.7167	0.6941	0.9273	0.6396	**0.9569**
gait	0.8287	**0.8400**	0.7759	0.8014	**0.9311**	0.5978	0.7710	0.6649	**0.7510**
ECG	0.8328	**0.8745**	0.8417	0.8169	**0.9500**	0.8577	0.9292	0.8167	**1.0000**
BP	0.8836	**0.8958**	0.7820	0.9622	**0.9722**	0.9336	0.9001	0.8225	**0.9492**

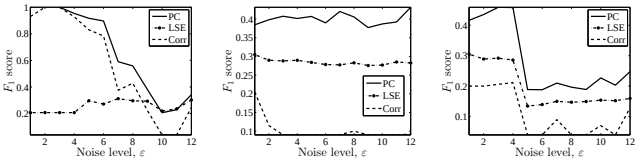
with a deterministic trend. These values for each data set from Table I are marked with vertical lines on Fig. 11 (d-g). Setting ϵ as shown on Fig. 11, we computed F_1, Precision and Recall values, that are listed in Table II.

As can be seen from Table II the proposed method performs better than the alternatives in terms of F_1 score and Precision in case of synthetic time series but fails to outperform both alternatives when applied to real data. Nevertheless, as Fig. 11 shows, if one chooses to permit less deviation from the ground truth $\mathcal{I}_{\mathcal{S}}{}^*$ and sets a smaller ϵ parameter, the proposed method will do better than alternatives, when applied to gait and blood pressure data sets. However, the quality of period extraction will drop for all three methods as well.

Table II indicates that the cross correlation estimation dominates its in terms of Recall, being less prone to miss segmentation points than its alternatives. On the contrary, it scores much worse in Precision. Similarly, the method based on least square estimation generally performs well in terms

of Precision, which means it makes less false positives, but demonstrates relatively low Recall. Compared to the alternatives, the proposed method scores evenly both Precision and Recall.

We have yet to define, why the alternatives demonstrate rather low quality on the synthetic set, but outperform the proposed method on the real data set. On one hand, we found that the proposed method performed better (in comparison the alternatives) when the structure of the data was more complicated. When period length T changes randomly with small variance, the alternatives, which extract periods of constant length, might work well enough. When T changes deterministically (see, for example Fig. 9), or varies greatly, the proposed method provides better results. However no evidence on dependance between relative performance on the noisiness level was found. Fig. 12 presents the dependencies of F_1-score on the noisiness level, measured as described in Section VI-A. Fig. 12 compares the proposed method (PC) and the alternatives (LSE and Corr). The proposed method outperforms the alternatives regardless of the noisiness level.



(a) Type 1 (b) Type 2 (c) Type 3

Fig. 12. The dependence of F_1-score on the noisiness (measured as variance of the period length) of synthetic data for the proposed method (PC) and the alternatives (LSE, Corr). The figure shows that as the noisiness increases, the proposed method outperforms the alternatives.

Nevertheless, the quality of period extraction drops for the proposed method as well with the increase of variance and complexity of periodicity. To improve performance of the proposed method in such cases, we plan to design a method of splitting the trajectory of periodic principal components with one cut instead of two cuts. Such procedure would make the method more precise for the time series of complex structures, because for such series the form of the trajectory is far from being a spiral we expect it to be. One loop of such trajectory often has self-intersections and internal sub-loops. Cutting such trajectory with symmetry axis leads to extracting abundant segmentation points (false positives). Our next goal is to design a robust "one-cut" method to reduce the number of false positives.

VIII. Conclusion

The paper discusses the problem of partitioning nearly periodic time series into period-like segments. Our motivation for solving this problem is to design the hierarchical framework for behavior analysis. Developing such framework would enable the creation and development of intellectual software that will support its users in maintaining healthy lifestyle monitoring the daily processes in order to predict unwanted changes in biometric characteristics. The framework can be used for applications connected with monitoring biosystems behavior. The usage of the framework with respect to the appropriate data will be helpful in defining the boundaries of normal and abnormal development of any biosystem as well distinguishing between types of process development and predicting the transition from normal to abnormal.

The framework deals with the data from body sensor network in bottom-up fashion. In case of accelerometry data, the lowest levels of hierarchy correspond to the time series of particular activity types and the highest levels describe sequences of activities. The features extracted at this level are processed to the next level, where activity recognition occurs. On yet higher levels of hierarchy the framework deals with sequences of activities to extract behavioral patterns. Different data sources yield different numbers of levels with their own interpretations. The purpose of this framework is to learn typical behaviors, and then apply the model to detect abnormal behavior. Solving period extraction problem contributes to the first step of this hierarchical model for the system of monitoring human health and behavior. Automatic partitioning of a wide range of time series into segments can be used for interpretable feature extraction. The key importance of this step is that the segmentation results are interpretable in the sense of the nature of the process: basically, we move from steps to activities and than to more complex sequences. A similar hierarchical approach is discussed in [35], [36], [37].

We claim that for the nearly periodic time series a fundamental period can be seen as interpretable segment. We propose a method of segmenting nearly periodical time series into fundamental periods. The ending points of the periods are defined through cutting the trajectory of a pair of principal components of the trajectory matrix. We describe a procedure of selecting a pair of principal components, corresponding to the fundamental periodic component of the time series. This method of period extraction is more general compared to those traditionally used as it does not rely on any assumptions the form of the signal. We compare our method to the two straightforward methods of period extraction, which are based on least squares estimation (or its asymptotical equivalent, periodogram maximization) and correlation analysis, on several real and synthetical data sets. Though the proposed method does not outperform the alternatives on all the data, it is at least comparable with the best-performing algorithm in all experiments.

Appendix
The proof of Theorem

A similar result was obtained in [30] for time series

$$X = \{x(1), \dots, x(i), \dots x(m)\}$$

of the form

$$x(i) = Ae^{\alpha i}\cos(2\pi w i + \phi)$$

with $w \in (0, 1/2)$, A, $\alpha \in \mathbb{R}$, $\phi \in [0, 2\pi)$ and $m \cdot w \in \mathbb{N}$. It was shown that if $\alpha \to 0$ with $m \to \infty$ so that $m \cdot \alpha \to \gamma \in \mathbb{R}$ the principal components $\mathbf{y}_1$ and $\mathbf{y}_2$ can be expressed as

$$y_1(l) = B_1 e^{\alpha l}\cos(2\pi w l + \phi_1),$$
$$y_2(l) = B_2 e^{\alpha l}\cos(2\pi w l + \phi_2),$$

$$\phi_1,\ \phi_2 \in [0, 2\pi),\ l = 1, \ldots, m - N + 1$$

and the difference $|\phi_1 - \phi_2| \to \pi/2$.

We consider more general form (5) of time series. According to [38], eigenvectors $\mathbf{v}_j$ of $\mathbf{H}$ form the orthonormal basis in the linear vector space defined by the columns of $\mathbf{H}$, thus for the time series of the form (5) we obtain [27] two eigenvectors $\mathbf{v}_1, \mathbf{v}_2$ with non-zero singular values $\lambda_1, \lambda_2 \neq 0$. The corresponding principal components $\mathbf{y}_1, \mathbf{y}_2$ have the form

$$y_1(l) = B_1(l)\cos(2\pi w l + \phi_1),$$
$$y_2(l) = B_2(l)\cos(2\pi w l + \phi_2).$$

Comparing the expression

$$\sum_{l=1}^{m-N+1} A_{i+l}A_{l+k}\cos(2\pi w(l+i)+\phi)\cos(2\pi w(l+k)+\phi)$$

for (i,k)-th element of $\mathbf{H}^{\mathsf{T}}\mathbf{H}$ with $N\mathbf{v}_i^{\mathsf{T}}\mathbf{\Lambda}\mathbf{v}_k$, we obtain (using the equation (4)) that

$$B_j(l) \sim A_l/\sqrt{N\lambda_l}.$$

Then $|A_i| \leq C \Rightarrow C_2 \leq B_1(l), B_2(l) \leq C_1$. The orthogonality condition $\mathbf{v}_1^{\mathsf{T}}\mathbf{v}_2 = 0$ yields

$$0 = \frac{1}{N}\sum_{l=1}^{m-N+1} B_1(l)^2\cos(2\pi w l+\phi_1)\cos(2\pi w l+\phi_2) =$$
$$\sum_{l=1}^{m-N+1} B_1(l)^2\cos(4\pi w l+\phi_1+\phi_2) \quad (11)$$
$$+\cos(\phi_1-\phi_2)\sum_{l=1}^{m-N+1} B^2(l). \quad (12)$$

Consider the first sum (Eq. (11)). Since $B_j(l) \leq C_1$,

$$\sum_{l=1}^{m-N+1} B_1(l)B_2(l)\cos(4\pi w l+\phi_1+\phi_2) \leq$$
$$C_1^2\sum_{l=1}^{m-N+1}\cos(4\pi w l+\phi_1+\phi_2) =$$
$$C_1^2\sum_{l=1}^{m-N+1}\cos(4\pi w\cdot l)\cos(\phi_1+\phi_2) -$$
$$C_1^2\sum_{l=1}^{m-N+1}\sin(4\pi w\cdot l)\sin(\phi_1+\phi_2).$$

For $\sum_l \cos(lx)$ and $\sum_l \sin(lx)$ we have:

$$\sum_{l=1}^{L}\cos(lx) = \frac{\cos x - 1 - \cos\big((L+1)x\big) + \cos Lx}{2-2\cos x},$$

$$\sum_{l=1}^{L}\sin(lx) = \frac{3\sin x - \sin\big((L+1)x\big) - \sin Lx}{2-2\cos x}.$$

Then, for any m we have the following upper constraint for the sum in (11):

$$\sum_{l=1}^{m-N+1} B_1(l)B_2(l)\cos(4\pi w l+\phi_1+\phi_2) \leq$$
$$C_1^2\left(\frac{2\cos(\phi_1+\phi_2)+5\sin(\phi_1+\phi_2)}{4}\right).$$

As for the sum in Eq. (12), since $A_i \geq C_2$, we obtain

$$\sum_{l=1}^{m-N+1} B^2(l) \geq (m-N+1)C_2^2.$$

If $\cos(\phi_1 - \phi_2) \not\to 0$, then the expression in Eq. (12) is unlimited. For the orthogonality condition to hold, the $|\phi_1 - \phi_2| \to \pi/2$ must hold as well.

That ends the proof. Let us now obtain the estimations of λ_1, λ_2 for the case when A_i is constant: $A_i \equiv A$. According to (4),

$$\lambda_j \approx \frac{A^2}{N}\sum_{l=1}^{N}\cos^2(2\pi w l) = \frac{A^2}{2N}\sum_{l=1}^{N} N(1+cos(4\pi w l))$$

or

$$\lambda_j \approx \frac{A^2}{N}\sum_{l=1}^{N}\sin^2(2\pi w l) = \frac{A^2}{2N}\sum_{l=1}^{N} N(1-cos(4\pi w l)).$$

Then the value of λ_j can be estimated as

$$\lambda_{\cos} \approx \frac{A^2}{2N}\int_0^N \big(1+\cos(4\pi w t)\big)dt = \frac{A^2}{2}\left(1-\frac{\sin(4\pi w N)}{4\pi w N}\right)$$
$$\text{for } y_j \sim \cos wj,$$
$$\lambda_{\sin} \approx \frac{A^2}{2N}\int_0^N \big(1-\cos(4\pi w t)\big)dt = \frac{A^2}{2}\left(1+\frac{\sin(4\pi w N)}{4\pi w N}\right)$$
$$\text{for } y_j \sim \sin wj.$$

The eigenvalues are asymptotically (with $N \to \infty$) equal. Moreover, when $w = 1/2$ and $T = 2$, we obtain only one principal component describing periodicity of X. This is why the method is limited to the cases of $w \in (0, 0.5]$.

REFERENCES

[1] L. Bao, S. Intille, "Activity recognition from user-annotated acceleration data", in. *Proceedings of the international conference on pervasive computing and communications*, pp. 1-7, 2004.

[2] J. Baek, G. Lee, W. Park, B.J. Yun, "Accelerometer signal processing for user activity detection", in *Knowledge-Based Intelligent Information and Engineering Systems, Lecture Notes in Computer Science*, vol. 3215, pp. 610-617, 2004.

[3] P.C. Petrantonakis, L.J. Hadjileontiadis, "Emotion Recognition From EEG Using Higher Order Crossings," *IEEE Transactions on Information Technology in Biomedicine*, vol.14, no.2, pp.186 197, March 2010.

[4] A. Grünerbl, A. Muaremi, V. Osmani, G. Bahle, S. hler, G. Trster, O. Mayora, C. Haring and P. Lukowicz, "Smart-Phone Based Recognition of States and State Changes in Bipolar Disorder Patients," *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 140-148, 2015.

[5] Xin Ma, Haibo Wang, Bingxia Xue, Mingang Zhou, Bing Ji, and Yibin Li, "Depth-Based Human Fall Detection via Shape Features and Improved Extreme Learning Machine," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 6, pp. 1915 - 1922, 2014.

[6] J. B. J. Bussmann, Y. M. van de Laar, M. P. Neeleman, H. J. Stam, "Ambulatory accelerometry to quantify motor behaviour in patients after failed back surgery: a validation study", *Pain* vol. 74, no. 23, pp. 15361, 1998.

[7] B. Najafi, K. Aminian, A. Paraschiv-Ionescu, F. Loew, C. J. Bula, P. Robert, "Ambulatory system for human motion analysis using a kinematic sensor: monitoring of daily physical activity in the elderly" in *IEEE Trans Biomed*, vol. 50, no. 6, pp. 711723, 2003.

[8] E. B. Mazomenos, D. Biswas, A. Acharyya, T. Chen, K. Maharatna, J. Rosengarten, J. Morgan, N. Curzen, "A Low-Complexity ECG Feature Extraction Algorithm for Mobile Healthcare Applications," *IEEE Journal of Biomedical and Health Informatics*, vol.17, no.2, pp.459 469, March 2013.

[9] Huaming Li, Jindong Tan, "Heartbeat-Driven Medium-Access Control for Body Sensor Networks," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 1, pp. 44-51, 2010.

[10] D. C. Mack, J. T. Patrie, P. M. Suratt, R. A. Felder, M. Alwan, "Development and Preliminary Validation of Heart Rate and Breathing Rate Detection Using a Passive, Ballistocardiography-Based Sleep Monitoring System," *IEEE Trans. Information Technology in Biomedicine*, vol.13, no.1, pp.111-120, Jan. 2009.

[11] A.M. Khan, Kyung Hee Young-Koo Lee, S. Y. Lee, Tae-Seong Kim, "A Triaxial Accelerometer-Based Physical-Activity Recognition via Augmented-Signal Features and a Hierarchical Recognizer," *IEEE Transactions on Information Technology in Biomedicine*, vol. 14, no. 5, pp. 1166 - 1172, 2010.

[12] J. R. Kwapisz, G. M. Weiss, S. A. Moore, "Activity Recognition using Cell Phone Accelerometers", in *Proceedings of the Fourth International Workshop on Knowledge Discovery from Sensor Data (at KDD-10)*, vol. 12, no. 2, pp. 74-82, 2010.

[13] J.M. Hausdorff et al. "Fractal dynamics of human gait: stability of long-range correlations in stride interval fluctuations". *Journal of Applied Physiology*, vol. 80, no. 5, pp. 1448-1457. 1996.

[14] A. M. Walker, "On the estimation of a harmonic component in a time series with stationary independent residuals", *Biometrika*, vol. 58, pp. 2136, 1971.

[15] E. J. Hannan, "The estimation of frequency", *Journal of Applied Probability*, vol. 10, pp. 510519, 1973.

[16] B. Boashash, "Estimating and interpreting the instantaneous frequency of a signalpart 2: algorithms and applications", in *Proceedings of the Institute of Electrical and Electronics Engineers (IEEE)*, vol. 80, no. 4, pp. 540-568, 1992.

[17] H. Kawahara, H. Katayose, A. de Cheveigne, R. D. Patterson, "Fixed point analysis of frequency to instantaneous frequency mapping for accurate estimation of f0 and periodicity", *EuroSpeech*, vol. 99, no. 6, pp. 2781-2784, 1999.

[18] G. Thuer, T. Verwimp, "Step Detection Algorithms for Accelerometers", PAPER OF THE E-LAB MASTER THESIS' 2008-2009, http://nitarc.be/map/paper/AMBIT_ThuerVerwimp.pdf.

[19] A. Brajdic, R. Harle. "Walk Detection and Step Counting on Unconstrained Smartphones", UbiComp 2013: Zurich, Switzerland, pp. 225-234, 2013.

[20] H. Ying, C. Silex, A. Schnitzer, S. Leonhardt, M.Schiek, "Automatic step detection in the accelerometer signal", *4th International Workshop on Wearable and Implantable Body Sensor Networks*, pp. 80-85, 2007.

[21] M. Nyan, F. Tay, K. Seah, Y. Sitoh, "Classification of gait patterns in the timefrequency domain", *Journal of Biomechanics*, vol. 39, no. 14, pp. 26472656, 2006.

[22] P. Barralon, N. Vuillerme, N. Noury, "Walk detection with a kinematic sensor: frequency and wavelet comparison", in *Proceedings of the 28th IEEE EMBS annual international conference*, vol. 1, pp. 1711-1714, 2006.

[23] A. Mannini, A. M. Sabatini, "Accelerometry-based classification of human activities using markov modeling", *Comp. Int. and Neurosc.*, 2011.

[24] S.J. Shultz et al. "Examination of muskoskeletal injuries". 2nd ed, North Carolina: Human Kinetics, 2005. p. 55-60.

[25] J. Loudon et al. "The clinical orthopedic assessment guide". 2nd ed. Kansas: Human Kinetics, 2008. p.395-408.

[26] André Gensler, Bernhard Sick, "Novel Criteria to Measure Performance of Time Series Segmentation Techniques". *In: T. Seidl, M. Hassani, C. Beecks (Eds.): Proceedings of the LWA 2014 Workshops: KDML, IR, FGWM, Aachen, Germany*, 8-10 September 2014.

[27] N. Golyandina, V. Nekrutkin and A. Zhigljavsky. "Analysis of Time Series Structure: SSA and related techniques". Chapman and Hall/CRC. 2001.

[28] A.N. Karkishchenko, V.B. Mnukhin, "Recovery of points symmetry in images of objects with reflectional symmetry", *Machine Learning and Data Analysis*, vol. 1, no. 5, pp. 621-631, 2013.

[29] A. N. Karkishchenko, V. B. Mnukhin. "Reflective Symmetrization of feature points in images". in *11th Conference (International) on Pattern Recognition and Image Analysis: New Information Technologies (PRIA-11-2013) Proceedings. Samara*, pp. 209-212. 2013.

[30] Th. Alexandrov, N. Golyandina. "Automatic extraction and forecast of time series cyclic components within the framework of SSA", in *Proceedings of the 5th Workshop on Simulation*, pp.45-50. 2005.

[31] P. P. Kanjilal, J. Bhattacharya and G. Saha. "Robust method for periodicity detection and characterization of irregular cyclical seriesin terms of embedded periodic components". *Physical review E*, vol. 59, no. 4, pp. 4013-4025. 1999.

[32] A. L. Goldberger et al, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals", *Circulation*, vol. 101, no. 23:e215-e220 [Circulation Electronic Pages; http://circ.ahajournals.org/cgi/content/full/101/23/e215], June 2000.

[33] T. Penzel, G. B. Moody, R. G. Mark, A. L. Goldberger, J. H. Peter, "The Apnea-ECG Database", *Computers in Cardiology*, vol. 27, pp. 255-258, 2000.

[34] Y. Ichimaru, G. B. Moody, "Development of the polysomnographic database on CD-ROM", *Psychiatry and Clinical Neurosciences*, vol. 53, pp. 175-177, April 1999.

[35] F. Zhou, F. De la Torre, J. K. Hodgins, "Hierarchical Aligned Cluster Analysis for Temporal Clustering of Human Motion", *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 35, no. 3, pp. 582-596, 2013.

[36] O. C. Jenkins, M. J. Mataric, "Deriving action and behavior primitives from human motion data", *In International Conference on Intelligent Robots and Systems*, pp. 2551-2556, 2002.

[37] P. Beaudoin, S. Coros, M. van de Panne, and P. Poulin, "Motion-motif graphs", in *ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, pp. 117-126, 2008.

[38] R. Vautard, M. Ghil, "Singular Spectrum Analysis in Nonlinear Dynamics with Applications to Paleoclimatic Time Series", *Physica D: Nonlinear Phenomena*, vol. 35, no. 3, pp. 395-424, 1989.