Shape anomaly detection under strong measurement noise: An analytical approach to adaptive thresholding

Alexander S. Krasichkov,[*] Eugene B. Grigoriev, and Mikhail I. Bogachev[†]
St. Petersburg Electrotechnical University, 5 Professor Popov Street, St. Petersburg 197376, Russia

Eugene M. Nifontov
Pavlov First Saint Petersburg State Medical University, 6/8 Leo Tolstoy Street, St. Petersburg 197022, Russia

(Received 23 July 2015; revised manuscript received 30 September 2015; published 29 October 2015)

We suggest an analytical approach to the adaptive thresholding in a shape anomaly detection problem. We find an analytical expression for the distribution of the cosine similarity score between a reference shape and an observational shape hindered by strong measurement noise that depends solely on the noise level and is independent of the particular shape analyzed. The analytical treatment is also confirmed by computer simulations and shows nearly perfect agreement. Using this analytical solution, we suggest an improved shape anomaly detection approach based on adaptive thresholding. We validate the noise robustness of our approach using typical shapes of normal and pathological electrocardiogram cycles hindered by additive white noise. We show explicitly that under high noise levels our approach considerably outperforms the conventional tactic that does not take into account variations in the noise level.

DOI: 10.1103/PhysRevE.92.042927 PACS number(s): 89.75.Kd, 87.18.Tt, 87.19.Hh, 87.85.Ng

I. INTRODUCTION

Detection and classification of shape anomalies is a ubiquitous problem arising in the analysis and interpretation of various experimental data. Prominent examples include pattern analysis and classification in such applications as microscopic imaging, mass spectrometry, nondestructive testing, and many others (see, e.g., [1–8], and references therein). Anomalous shape detection is also essential in the analysis of time series exhibiting periodic patterns superimposed by irregular stochastic fluctuations. Common examples include detection of significant changes in the shapes of complex vital signals such as an electrocardiogram (ECG) or electroencephalogram exhibiting quasiperiodic variability [9] as well as finding deviations of certain physiological parameters from typical patterns, e.g., the circadian blood pressure profile that is different from typical for a given individual [10–12]. More applications of atypical shape detection could be found in the analysis of short-term cycles in atmospheric, climatic, hydrologic, geomagnetic, and other geophysical data sets [13–16]. Recently, several sophisticated methodologies to handle shape analysis and classification in quasiperiodic time series have been suggested, including dynamic time warping [17,18], a complex-network-based approach [19,20], the Bayesian framework-based approach [21], dimensionality reduction techniques [22], and some others. While demonstrating improved performance and often providing additional information for the shape anomaly classification, many of these techniques require considerable computational efforts as well as initial data accumulation. In contrast, online monitoring systems such as ECG analyzers often require an immediate reaction to the shape anomaly occurrence by algorithms that are able to run on low-performance wearable devices over a long time. Therefore, despite the drastic increase in the performance of wearable devices in recent years, there is still a demand for simple online shape anomaly detection methods.

The most straightforward approach to either detection or classification of various shape anomalies in the observational data is performed by its comparison to either single or multiple reference shapes using a certain similarity score. In the detection problem the similarity score is next compared to a certain threshold that provides the level of dissimilarity that should be treated as an anomaly, while in the classification problem the shape with the highest similarity score is selected. Widely used similarity scores are commonly based on the variants of the inner product such as the cosine similarity score, cross covariance, or the cross-correlation coefficient [19,23–28]. In the experimental data, moderate shape anomalies can be either imitated or hindered by the measurement noise. Accordingly, the reduction in the similarity score can be caused either by the true shape anomaly or by the increase in the noise level as well as by the combination of these effects. A prominent example is the long-term ECG analysis where the measurement noise level can vary rapidly and drastically with the changes in the muscle activity [29]. Therefore, variations in the noise level require considerable adjustments of the decision threshold.

In this paper we suggest an analytical approach to the threshold adjustment with changes in the noise level. We note that the cosine similarity score is equivalent to the correlation coefficient for data sets with zero mean value. Since both the mean value and slow observational data variations (like the baseline drift in the observational ECG) can be easily eliminated as a preliminary step to prepare the data for further analysis, we next focus on the cosine similarity score and additive white noise model [30,31]. We compare our analytical treatment with the computer simulation results and show that there is a perfect match. Using several typical ECG shapes as prominent examples, we also provide the characteristics of the efficiency and noise robustness of the adaptive thresholding in

[*] Also at: Federal North-West Medical Research Centre, 2 Akkuratova Street, St. Petersburg 197341, Russia.

[†] Corresponding author: rogex@yandex.com

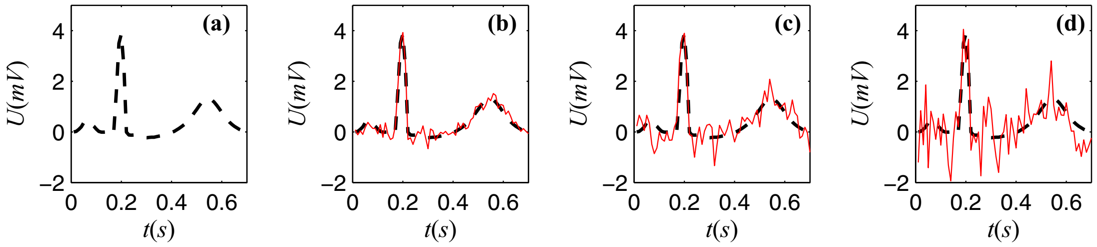

FIG. 1. (Color online) (a) Noise-free sample shape of an ECG cycle corresponding to a normal heartbeat (dashed line). (b)–(d) Same ECG cycle (dashed line) with different levels of additive Gaussian white noise (solid lines): (b) $h = 7.46 \times 10^{-4}$, (c) $h = 4.61 \times 10^{-3}$, and (d) $h = 1.19 \times 10^{-2}$.

comparison with the fixed thresholding under significant noise level variations.

II. ANALYTICAL TREATMENT

Let us consider a complex shape of a typical normal ECG cycle shown in Fig. 1(a) as a reference signal $S_{\rm ref}(i)$. The observational signal $S(i)$ is hindered by additive Gaussian white noise $n(i)$ with zero mean value and σ_n^2 variance. Both signals are sampled at time points $i = 1,\dots,N$. Then the cosine similarity score r between $S_{\rm ref}(i)$ and $S(i)+n(i)$ is given by

$$\begin{aligned} r &= \frac{\sum_{i=1}^N S_{\rm ref}(i)[S(i)+n(i)]}{\sqrt{\sum_{i=1}^N S_{\rm ref}^2(i)\sum_{i=1}^N [S(i)+n(i)]^2}} \\ &= \frac{\rho + \sum_{i=1}^N \left[S_{\rm ref}(i)/\sqrt{E_{S_{\rm ref}}}\right][n(i)/\sqrt{E_S}]}{\sqrt{\sum_{i=1}^N [S(i)/\sqrt{E_S}+n(i)/\sqrt{E_S}]^2}} \\ &= \frac{\rho + \sum_{i=1}^N x(i)\Delta(i)}{\sqrt{1+2\sum_{i=1}^N y(i)\Delta(i)+\sum_{i=1}^N \Delta^2(i)}}, \end{aligned} \qquad (1)$$

where $E_{S_{\rm ref}} = \sum_{i=1}^N S_{\rm ref}^2(i)$ is the energy of the reference signal, $E_S = \sum_{i=1}^N S^2(i)$ is the energy of the (assumed noise-free) observational signal, ρ is the cosine similarity score between the reference signal and the (noise-free) observational signal, $x(i) = \frac{S_{\rm ref}(i)}{\sqrt{E_{S_{\rm ref}}}}$ is the normalized reference signal, $y(i) = \frac{S(i)}{\sqrt{E_S}}$ is the normalized (noise-free) observational signal, $\Delta(i) = \frac{n(i)}{\sqrt{E_S}}$ is the normalized noise with variance $\sigma^2 = \frac{\sigma_n^2}{E_S}$, and $\Delta(i) = N(0,\sigma)$.

Obviously, r will decrease with increasing the noise level, even when $S_{\rm ref}(i)$ and $S(i)$ exhibit identical shapes. For an approximate calculation of r, we next replace the radicand in the denominator of Eq. (1) with

$$\xi = 1 + 2\sum_{i=1}^N y(i)\Delta(i) + \sum_{i=1}^N \Delta^2(i) \qquad (2)$$

and focus on the expansion of $\frac{1}{\sqrt{\xi}}$ by the nth-order polynomial

$$\frac{1}{\sqrt{\xi}} \approx d\xi^n + \cdots + a\xi^2 + b\xi^1 + c. \qquad (3)$$

To find the polynomial coefficients, let us first consider the typical range of the random variable ξ. The second term in the expression (2) is a Gaussian random variable and the third term is approximately Gaussian for $N > 30$ [32]. Thus we next consider ξ as a Gaussian random variable with mean $m(\xi/h) = hN + 1$ and variance $D(\xi/h) = 2h(hN+2)$, where

$$h = \sigma^2 = \frac{\sigma_n^2}{E_S}$$

is a parameter characterizing current noise level.

Next one can determine the probability p that the random variable ξ fits within a given range. For example, the probability that ξ deviates from its mean value by more than two of its standard deviations is given by

$$p(|\xi - m\{\xi/h\}| > 2\sqrt{D\{\xi/h\}}) < 0.05,$$

while the deviation range is given by

$$\xi_l(h) = 2\sqrt{2h(hN+2)}. \qquad (4)$$

Equation (4) shows that this range depends on the number of data points N and on the noise level parameter h. In the following we define the range using (4). The range of typical ξ values is given by $[m\{\xi/h\} - \xi_l(h), m\{\xi/h\} + \xi_l(h)]$ and the condition $m(\xi/h) - \xi_l(h) > 0$ has to be fulfilled.

With the knowledge of the range for $\xi_l(h)$, one can find the approximation coefficients $d,\dots,a,b,c$ for a given h and $N[(\xi_l) = \xi_l(h), m(\xi) = m(\xi/h)]$ using the least-mean-squares algorithm as the minimum of the function

$$\begin{aligned} &f(m\{\xi\},\xi_l,d,\dots,a,b,c) \\ &\quad = \int_{m\{\xi\}-\xi_l}^{m\{\xi\}+\xi_l} \left(\frac{1}{\sqrt{\xi}} - d\xi^n - \cdots - a\xi^2 - b\xi^1 - c\right)^2 d\xi. \end{aligned} \qquad (5)$$

At a first approximation we next reduce to the polynomials of the first and the second order. The minimum of expression (5) for the second-order polynomial will be achieved at

$$\begin{aligned} c &= \frac{1}{2\xi_l}(E - aC - bD), \\ a &= \left[b\left(\frac{DC}{2\xi_l} - B\right) + P - \frac{EC}{2\xi_l}\right]\frac{2\xi_l}{2\xi_l A - C^2}, \\ b &= \frac{(2\xi_l T - ED)(2\xi_l A - C^2) - (2\xi_l B - CD)(2\xi_l P - EC)}{(2\xi_l C - D^2)(2\xi_l A - C^2) - (2\xi_l B - CD)^2} \end{aligned} \qquad (6)$$

TABLE I. Statistical characteristics of the random variable ξ and of the first- and second-order polynomial approximation coefficients for the sample shapes shown in Figs. 1(b)–1(d).

			Second-order polynomial			First-order polynomial	
h	$m\{\xi\}$	$D\{\xi\}$	a	b	c	b	c
7.46×10^{-4}	1.052	0.055	0.331	−1.162	1.831	−0.464	1.464
4.61×10^{-3}	1.326	0.147	0.189	−0.837	1.644	−0.322	1.314
1.19×10^{-2}	1.836	0.260	0.086	−0.522	1.407	−0.206	1.124

and for the first-order polynomial at

$$b = \frac{2\xi_l T - ED}{2\xi_l C - D^2},$$
$$c = \frac{1}{2\xi_l}\left(E - \frac{2\xi_l T - ED}{2\xi_l C - D^2}\right), \tag{7}$$

where $A = \int_{m\{\xi\}-\xi_l}^{m\{\xi\}+\xi_l} \xi^4 d\xi$, $B = \int_{m\{\xi\}-\xi_l}^{m\{\xi\}+\xi_l} \xi^3 d\xi$, $C = \int_{m\{\xi\}-\xi_l}^{m\{\xi\}+\xi_l} \xi^2 d\xi$, $D = \int_{m\{\xi\}-\xi_l}^{m\{\xi\}+\xi_l} \xi d\xi$, $E = \int_{m\{\xi\}-\xi_l}^{m\{\xi\}+\xi_l} \frac{1}{\sqrt{\xi}} d\xi$, $T = \int_{m\{\xi\}-\xi_l}^{m\{\xi\}+\xi_l} \frac{\xi}{\sqrt{\xi}} d\xi$, and $P = \int_{m\{\xi\}-\xi_l}^{m\{\xi\}+\xi_l} \frac{\xi^2}{\sqrt{\xi}} d\xi$.

Remarkably, expressions (5)–(7) indicate that the polynomial coefficients depend only on the parameters h and N and thus are independent from the shape itself. Thus, in the following we denote a, b, and c by $a_{h,N}$, $b_{h,N}$, and $c_{h,N}$, respectively.

Table I summarizes the statistical characteristics of the random variable ξ and of the polynomial coefficients for the first- and second-order approximations for the sample shape consisting of $N = 70$ data points shown in Fig. 1(a) calculated for $h = 7.46 \times 10^{-4}$ [Fig. 1(b)], $h = 4.61 \times 10^{-3}$ [Fig. 1(c)], and $h = 1.19 \times 10^{-2}$ [Fig. 1(d)]. Figure 2 shows the approximation of $1/\sqrt{\xi}$ as a function of ξ exemplified for the noisy sample shape shown in Fig. 1(d) ($h = 1{,}19 \times 10^{-2}$). The solid line is the $1/\sqrt{\xi}$ function, the triangles are its second-order polynomial approximation, and the circles are its first-order polynomial approximation. The approximation error (defined as the normalized maximum absolute value deviation of the approximation function from the original function) for the second-order polynomial fitted on the interval $[m\{\xi\} - \xi_l, m\{\xi\} + \xi_l]$ is below 0.01% and for the first-order polynomial it is below 1%.

Since r is given by (1), an approximation of ξ by the first-order polynomial leads to the ansatz

$$r_1 = \left(\rho + \sum_{i=1}^{N} x(i)\Delta(i)\right)\left[b_{h,N}\left(1 + 2\sum_{i=1}^{N} y(i)\Delta(i) + \sum_{i=1}^{N} \Delta^2(i)\right) + c_{h,N}\right]. \tag{8}$$

Simple calculations lead to the expression for the mean value of the random variable r_1,

$$\begin{aligned} m\{r_1\} = \overline{r} &= \overline{\left(\rho + \sum_{i=1}^{N} x(i)\Delta(i)\right)\left[b_{h,N}\left(1 + 2\sum_{i=1}^{N} y(i)\Delta(i) + \sum_{i=1}^{N} \Delta^2(i)\right) + c_{h,N}\right]} \\ &= \rho\left[b_{h,N}\left(1 + 2\overline{\sum_{i=1}^{N} y(i)\Delta(i)} + \overline{\sum_{i=1}^{N} \Delta^2(i)}\right) + c_{h,N}\right] + b_{h,N}\overline{\sum_{i=1}^{N} x(i)\Delta(i)} \\ &\quad + 2b_{h,N}\overline{\sum_{i=1}^{N} x(i)\Delta(i)\sum_{i=1}^{N} y(i)\Delta(i)} + b_{h,N}\overline{\sum_{i=1}^{N} x(i)\Delta(i)\sum_{i=1}^{N} \Delta^2(i)} + c_{h,N}\overline{\sum_{i=1}^{N} x(i)\Delta(i)} \\ &= \rho[b_{h,N}(1 + hN) + c_{h,N}] + 2b_{h,N}h\sum_{i=1}^{N} x(i)y(i) \\ &= \rho[b_{h,N}(1 + hN) + c_{h,N}] + 2b_{h,N}h\rho = \rho[b_{h,N}(h(N + 2) + 1) + c_{h,N}], \end{aligned} \tag{9}$$

where the overline denotes averaging. For $N \gg 1$ the expression (9) can be replaced by $m(r_1) = \rho[b_{h,N}(hN + 1) + c_{h,N}]$. Again, the expression (9) shows that the mean value depends only on h and N but not on the shape itself.

By using the second-order polynomial approximation of ξ one obtains

$$r_2 = \left(\rho + \sum_{i=1}^{N} x(i)\Delta(i)\right)\left[a_{h,N}\left(\sum_{i=1}^{N} [y(i) + \Delta(i)]^2\right)^2 + b_{h,N}\left(\sum_{i=1}^{N} [y(i) + \Delta(i)]^2\right) + c_{h,N}\right]. \tag{10}$$

In this case, after some routine calculations, the mean of the random variable r_2 can be expressed as

$$m(r_2) = a_{h,N}\{h^2[N^2 + 2N + 4\rho(N + 2)] + \rho h(2N + 8) + \rho\} + \rho\{b_{h,N}[h(N + 2) + 1] + c_{h,N}\}.$$

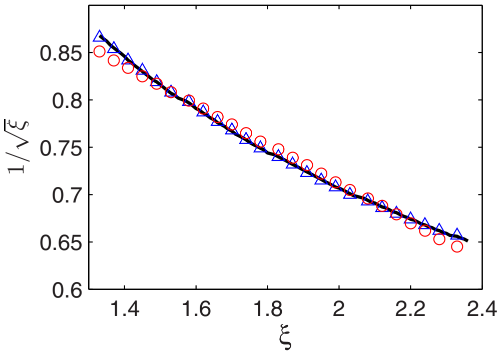

FIG. 2. (Color online) Approximation of the denominator of (1) exemplified for the sample ECG cycle shape with additive Gaussian noise [shown in Fig. 1(d)]. The solid line is the $1/\sqrt{\xi}$ function, while the circles and the triangles are its first- and second-order approximations, respectively.

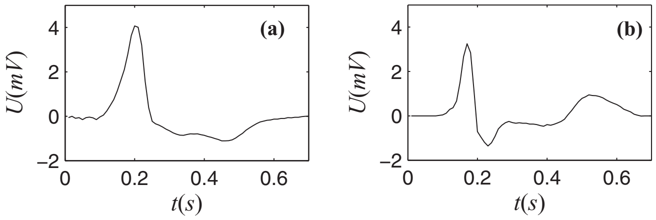

FIG. 4. Typical shapes corresponding to two single ECG cycles with different pathologies.

For $N \gg 1$ the approximations $2N+8 \approx 2N$, $N+2 \approx N$, and $N^2+2N+4\rho(N+2) \approx N^2$ lead to the ansatz

$$\begin{aligned} m(r_2) \approx\ & a_{h,N}[\rho + h^2(N^2+\rho 2N)] \\ & + \rho[b_{h,N}(hN+1) + c_{h,N}]. \end{aligned} \tag{11}$$

Once the reference and the (noise-free) observational shapes $S_{\text{ref}}(i)$ and $S(i)$ are identical ($\rho = 1$), one next obtains

$$m(r_2) \approx a_{h,N}(hN+1)^2 + b_{h,N}(hN+1) + c_{h,N}.$$

Figure 3 shows the dependence of the mean value on the noise level parameter h for the identical reference and observational shapes ($\rho = 1$), now corresponding to the shape of a typical pathological ECG cycle displayed in Fig. 4(a). The triangles show the results of computer simulations (20 000 iterations), while the solid lines show the approximations by the first-order polynomial [Fig. 3(a)] and by the second-order polynomial [Fig. 3(b)].

Based on the series of simulations for various reference and analyzed observational shapes corresponding to various ECG modifications that are typical for ventricular arrhythmia [shown in Fig. 4(a)] and other disorders such as myocardial infarction (not shown), we conclude that in most cases the first-order polynomial (8), or simply a linear approximation, is sufficiently accurate. While we realize that in some cases a second-order polynomial approximation (10) may be required, in the following example we find that the linear approximation (8) is sufficiently accurate and thus employ it to find the variance of the random variable r obtaining

$$\begin{aligned} M_2(r_1) =\ & h^3 b_{h,N}^2 (N^2+6N+8) \\ & + h^2\big[2b_{h,N}(b_{h,N}+c_{h,N})(N+2) + 4b_{h,N}^2(2\rho^2+1) \\ & + 8b_{h,N}^2\rho^2(6-N)\big] + h\big[(b_{h,N}+c_{h,N})^2 \\ & + 4b_{h,N}\rho^2(b_{h,N}+c_{h,N}) + 4b_{h,N}^2\rho^2\big]. \end{aligned}$$

For $N \gg 1$, since $N+2 \approx N$, $6-N \approx -N$, and $N^2+6N+8 \approx N^2+6N$, the above expression can be replaced by

$$\begin{aligned} M_2(r) \approx\ & b_{h,N}^2 h^3(N^2+6N) + \big[b_{h,N}^2(6\rho^2+2) \\ & + 2b_{h,N}c_{h,N}\big]h^2N + \big[b_{h,N}^2(8\rho^2+1) \\ & + 2b_{h,N}c_{h,N}(1+2\rho^2) + c_{h,N}^2\big]h. \end{aligned} \tag{12}$$

For $\rho = 1$ this expression can be further simplified

$$\begin{aligned} M_2(r_1) \approx\ & b_{h,N}^2 h^3(N^2+6N) + \big(8b_{h,N}^2 + 2b_{h,N}c_{h,N}\big)h^2N \\ & + (3b_{h,N}+c_{h,N})^2 h. \end{aligned} \tag{13}$$

Figure 5 shows the dependence of the variance of the random variable r on the noise level parameter h for the identical ($\rho = 1$) reference and observational shapes [shown in Fig. 4(a)] for $N = 70$. The triangles correspond to the computer simulation data (20 000 iterations), while the solid

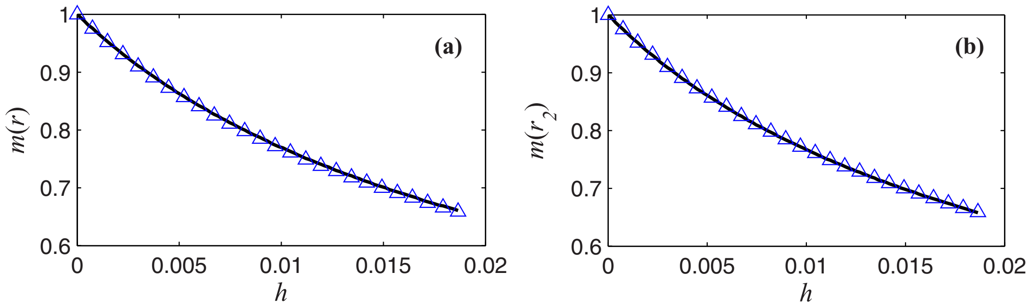

FIG. 3. (Color online) Dependence of the mean value of the cosine similarity score r on the noise level parameter h for the identical reference and observational shapes ($\rho = 1$), in particular, the typical shape of a single pathological ECG cycle, displayed in Fig. 4(a). The triangles correspond to the results of computer simulations (20 000 iterations), while the solid line shows the approximations by (a) the first-order polynomial [see Eq. (9)] and (b) the second-order polynomial [see Eq. (11)].

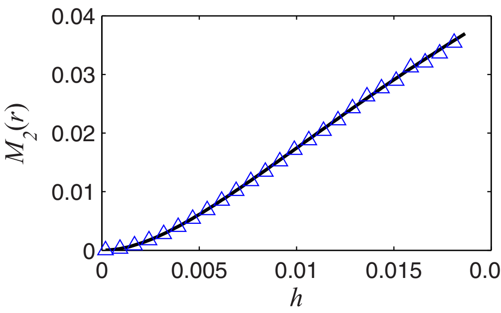

FIG. 5. (Color online) Dependence of the variance of the cosine similarity score r on the noise level parameter h for the identical ($\rho = 1$) reference and observational shapes [shown in Fig. 4(a)] at $N = 70$. The triangles correspond to the computer simulation results (20 000 iterations), while the solid line shows the analytical treatment [see Eq. (13)].

line shows the analytical treatment (13). The figure shows that the analytical approximation of the variance and the results of computer simulations are in good agreement.

Next we focus on the estimate of the correlation coefficient r^* between the random variables r_1 (8) and $\eta = \sum_{i=1}^{N} \Delta^2(i)$ [33]. The mean and the variance of r^* are given by

$$m(\eta) = hN, \quad M_2(\eta) = 2h^2N.$$

Then taking into account (9), the covariance function between r and η is given by

$$K(r,\eta) = 2b_{h,N}\rho h^2(N+2) \approx 2b_{h,N}\rho h^2 N.$$

Therefore, the correlation coefficient r^* can be estimated as

$$r^* = \frac{b_{h,N}\rho\sqrt{2hN}}{\sqrt{b_{h,N}^2 h^2(N^2+6N) + \left(8b_{h,N}^2 + 2b_{h,N}c_{h,N}\right)hN + (3b_{h,N}+c_{h,N})^2}}. \tag{14}$$

Figure 6 shows the dependence of r^* on the noise level parameter h for the reference and observational shapes of the same type shown in Fig. 1 ($\rho = 1$ and $N = 70$). The triangles were obtained by computer simulations (20 000 iterations), while the solid line shows an analytical treatment according to (14). The figure shows that for $h \in [0; 3.45 \times 10^{-3}]$, r and η exhibit pronounced negative correlation $r^* < -0.8$. Therefore, at a first approximation we consider $r \approx -\frac{M_2(r)}{M_2(\eta)}[\eta - m(\eta) + m(r)]$. Expression (14) can be fitted by the Gaussian distribution with the probability density function $W(r)$. In a more general case, for an arbitrary r^*, a certain approximation can be given by an Edgeworth expansion [34]. As the null hypothesis we assume that the empirical distribution is described by the first three terms of the Edgeworth expansion. To test this, we used the χ^2 test and the Kolmogorov-Smirnov test. The returned $p = 0.22$ in the χ^2 test and $p = 0.10$ in the Kolmogorov-Smirnov test indicate that both tests fail to reject the null hypothesis with 95% confidence level.

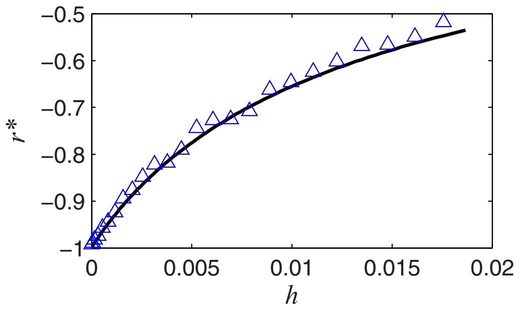

FIG. 6. (Color online) Dependence of the correlation coefficient estimate r^* on the noise parameters h for the identical reference and observational shapes shown in Fig. 1(a) ($\rho = 1$ and $N = 70$). The triangles were obtained by computer simulations (20 000 iterations), while the solid line shows the analytical treatment according to Eq. (14).

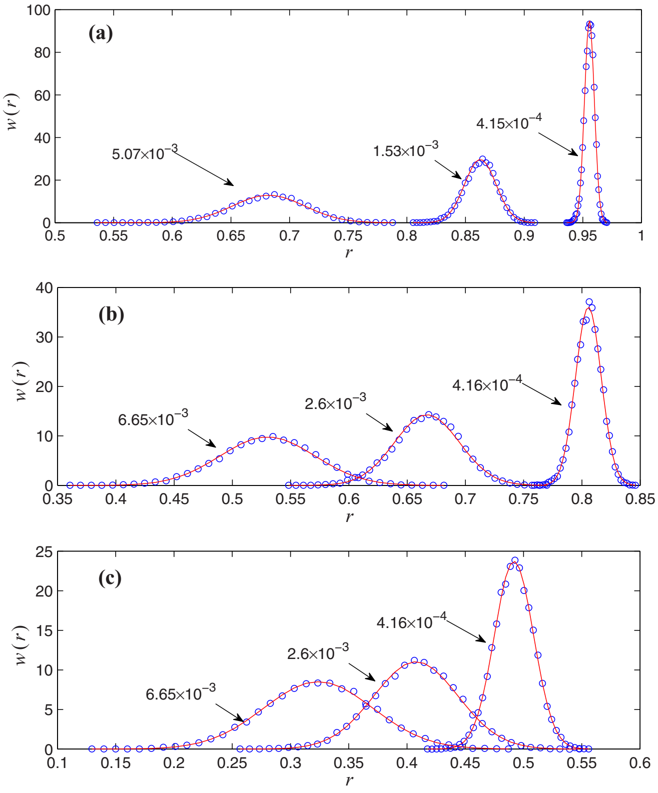

FIG. 7. (Color online) Probability distribution functions $w(r)$ of the cosine similarity score r for different noise parameters h (annotated by arrows in the figure) for (a) the same reference and analyzed ECG shapes shown in Fig. 1(a) ($\rho = 1$ and $N = 70$), (b) the reference shape shown in Fig. 1(a) and analyzed shape shown in Fig. 4(b) ($\rho = 0.84$ and $N = 70$), and (c) the reference shape shown in Fig. 1(a) and analyzed shape shown in Fig. 4(a) ($\rho = 0.505$ and $N = 70$). Circles correspond to the computer simulation results (20 000 iterations), while the solid lines show their approximations by a Gaussian distribution.

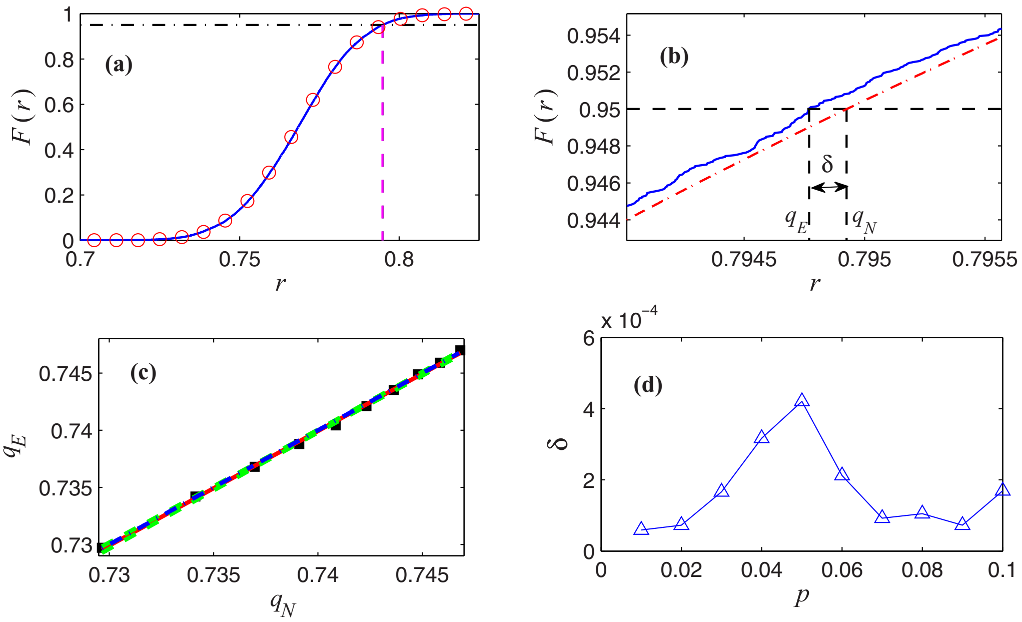

FIG. 8. (Color online) (a) Comparison of the quantiles from the empirical cumulative distribution function (CDF) $F(r)$ of the cosine similarity score r (q_E, solid line) and the Gaussian CDF (q_N, circles), respectively. The horizontal dash-dotted line corresponds to the chosen false negative rate (equal to 0.05 in this example). The vertical dashed lines show the determined quantile values. (b) Same as in (a) but the plot is zoomed in on the area of the intersection of lines. The solid line corresponds to the empirical CDF of the cosine similarity score r, while the dash-dotted line shows the Gaussian CDF. The difference between the quantiles is defined as $\delta = |q_N - q_E|$. (c) Linear regression line between the quantiles q_N and q_E (solid line), with the squares corresponding to the sample data, dash-dotted line showing the $q_E = q_N$ line, and the dashed line corresponding to the 95% confidence interval (all lines are very close to each other). (d) Dependence of the difference between the quantiles δ on the false negative rate p.

III. ADAPTIVE THRESHOLDING EFFICIENCY

Figure 7 shows the distribution of the cosine similarity score r as a function of h [Fig. 7(a)] for the identical reference and observational shapes [$\rho = 1$; see Fig. 1(a)] as well as for different reference and observational shapes characterized by Fig. 7(b) [$\rho = 0.84$; see Fig. 4(b)] and Fig. 7(c) [$\rho = 0.505$; see Fig. 4(a)]. Circles correspond to the computer simulation results (20 000 iterations), while solid lines show approximations by a Gaussian distribution. As one can see from Fig. 7, the Gaussian distribution is almost identical to the empirical distribution. Thus we next check whether the second and third terms in the Edgeworth expansion could be ignored. This leads to a simple Gaussian distribution that makes it easier to find the quantiles. To further test whether such a substitution by a Gaussian distribution is legitimate, we next compare the quantile estimate q_E obtained from the empirical cosine similarity score distribution and a similar estimate q_N from the respective Gaussian approximation [shown by open symbols and a solid line in Fig. 8(a), respectively]. While visual inspection indicates data collapse, the discrepancy can be observed in a zoomed plot [see Fig. 8(b)]. We next test the accuracy of the approximation for various thresholds that correspond to the quantiles between 0.1 and 0.01 by linear regression analysis [the corresponding quantile-quantile plot with the linear regression fit and its confidence intervals are shown in Fig. 8(c)] and confirm that very good agreement can be observed in all studied cases. Finally, Fig. 8(d) shows that the relative quantile estimation error is about three orders of magnitude below the threshold level for all ten quantiles studied. Thus it seems justified to conclude that the Gaussian approximation is legitimate for the cases studied.

The above example has a straightforward practical implication, provided the reference shape corresponds to the normal ECG cycle, while the observational shapes correspond to two different pathological ECG cycles. Accordingly we suggest that in the shape anomaly detection algorithm the decision threshold level should be adaptively adjusted with the changes in the noise level. The decision threshold can be defined as a quantile of the analytical distribution that corresponds either to a fixed false positive rate or to a fixed false negative rate. In life-threatening issues such as online ECG analysis where the cost a single error is high, the false negative rate is usually fixed. In our case, since the final distributions can be well approximated by Gaussian distributions (as shown above), instead of estimating the quantile numerically, one can calculate the threshold from the mean and the standard deviation of the final distribution. In particular, according to Eqs. (9) and (12), one obtains

$$q_{\text{adapt}} = m(r) - x\sqrt{M_2(r)},$$

where x is a fixed prefactor that depends on the chosen false negative rate.

Figure 9 exemplifies the adaptive thresholding for four different noise levels. For weak noise conditions [see Figs. 9(a) and 9(b)], when the distributions do not overlap, finding a fixed optimized threshold is trivial. However, if the noise level changes with time, a fixed threshold that appears perfect for

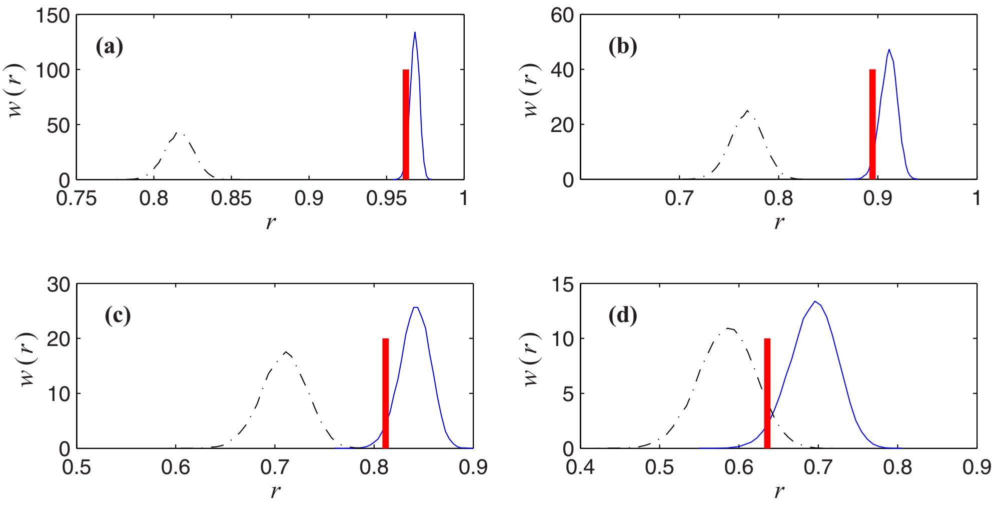

FIG. 9. (Color online) Probability density functions $w(r)$ of the cosine similarity scores r for (a) $h = 2.92 \times 10^{-4}$, (b) 8.95×10^{-4}, (c) 0.0018, and (d) 0.0047. The solid line corresponds to the distribution of the cosine similarity between the same reference and analyzed shapes shown in Fig. 1(a), while the dash-dotted line corresponds to the normal reference [Fig. 1(a)] and the pathological observational [Fig. 4(b)] shapes. Vertical lines exemplify the adaptive threshold q_{adapt} corresponding to the fixed false negative rate.

Fig. 9(a) (e.g., 0.9) becomes less and less adequate when the noise level increases [see Figs. 9(b)–9(d)]. In contrast, the suggested adaptive thresholding procedure guarantees the fixed false negative rate and this way outperforms the fixed thresholding tactic under strong noise conditions. Of course, the implementation of this solution in practical issues requires an online noise level estimation algorithm to be running to feed the decision making algorithm with the current noise level.

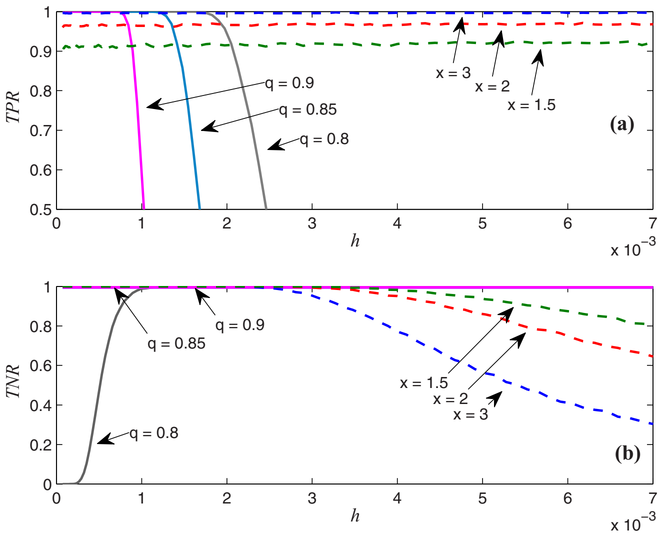

FIG. 10. (Color online) (a) True positive and (b) true negative rates as a function of the noise level parameter h. Solid lines refer to the constant thresholding q, while dashed lines refer to the suggested adaptive thresholding algorithm. The reference signal is always given by the same normal ECG shape shown in Fig. 1(a), while the observational signal is given (with equal probability) either by the same normal ($\rho = 1$) or by the pathological ECG shape shown in Fig. 4(b) with a cross-cosine similarity score of $\rho = 0.84$ between them.

Finally, we analyzed the noise robustness of the suggested approach in comparison with the fixed threshold tactic. Figure 10(a) shows the true positive rate (TPR) as a function of the noise level parameter h for three different $x = 1.5$, 2, and 3. The solid lines refer to three different fixed thresholds q, while the dashed lines correspond to the suggested adaptive thresholding procedure. Figure 10(a) illustrates the case where the reference shape is given by the normal ECG and the observational shape is given either by the normal or by the pathological ECG with cross-cosine similarity scores between them $\rho = 1$ and 0.84, respectively. The figure shows that the true positive rate decreases with the increase of the noise level parameter h. In contrast, when using an adaptive threshold (dashed lines) the false negative rate is fixed and can be determined analytically from Eqs. (9) and (12). Thus, by using the adaptive threshold, it is possible to achieve a constant true positive rate. Figure 10(b) shows the true negative rate (TNR) as a function of the noise level parameter h. The figure shows that by using the suggested approach with adaptive thresholding, a higher shape detection performance can be achieved under high noise levels.

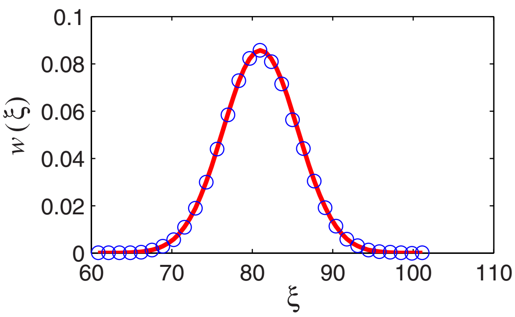

FIG. 11. (Color online) Empirical distribution of Eq. (2) (circles) and the approximation of this distribution by a Gaussian distribution (solid line) for the uniform noise scenario.

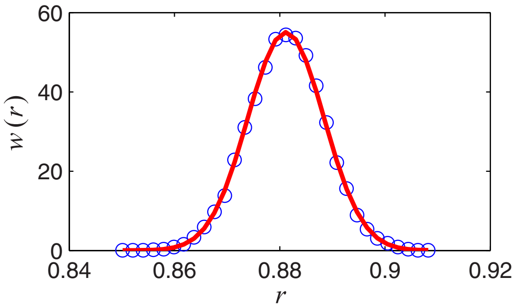

FIG. 12. (Color online) Distribution of the cosine similarity score r for the uniform noise scenario. The circles show the empirical distribution while the solid line corresponds to the approximation by the Gaussian distribution.

We note that our approach can be straightforwardly extended to the non-Gaussian noise case. According to the central limit theorem, the radicand in the denominator of Eq. (1) ξ can be described by a Gaussian distribution for large N. As an example, for the uniform distribution we obtain the probability distribution of ξ shown in Fig. 11. The circles show the empirical distribution of ξ, while the solid line corresponds to its Gaussian approximation. By using the Kolmogorov-Smirnov test, we obtained $p = 0.28$, indicating that the null hypothesis that the empirical data can be described by the Gaussian distribution cannot be rejected with 95% confidence level. Carrying out similar transformations, we can obtain analogs of the expressions (9)–(12) in the case of the uniform distribution. The distribution of the cosine similarity score r in the case of non-Gaussian noise according to the central limit theorem can also be described by a Gaussian distribution (see Fig. 12), as we have also confirmed empirically by a Kolmogorov-Smirnov test ($p = 0.22$).

IV. CONCLUSION

In summary, we have obtained an analytical solution of the cosine similarity score distribution between a given reference shape and an observational shape hindered by additive white noise. Shape similarity measures such as the cosine similarity score and the correlation coefficient are common in many applications, especially in those where the analyzed signals exhibit pronounced amplitude variations with time, since these measures are not sensitive to such variations. We also found analytical expressions for the mean and the variance of the cosine similarity score. We further confirmed our theoretical calculations by computer simulations and found a nearly perfect match. As a practical example, we considered shapes that correspond to either healthy or pathological ECG cycles. We have shown explicitly that our findings lead to the enhancement of the pathological ECG shape detection accuracy that can be applied to further improvements in the automatic ECG analysis algorithms. We think that our findings might also be helpful in detecting atypical patterns in short-term cycles of various physiological quantities, either directly measured such as heart rate or blood pressure [11,12] or derived from noisy observational records such as baroreflex sensitivity [35], that might be indicators of serious physiological disorders. They might also be useful in detecting anomalies of short-term cyclic patterns in atmospheric, climatic, hydrological, geomagnetic, and other geophysical data sets exhibiting pronounced quasiperiodic cycles. However, we note that many of these data sets exhibit long-term correlations [36–39] that are superimposed to quasiperiodic oscillations and thus linear and nonlinear correlations in their fluctuations should be additionally taken into account.

The proposed approach may also be used as part of a more complex problem, for example, in the construction of a network where the nodes are characterized by the noisy observational quasiperiodic data sources and the links between the nodes are considered as significant once the cross correlation between them exceeds a given threshold [40]. Very recently, network-based solutions found a number of successful applications in various quasiperiodic data analysis and event prediction such as sleep stages recognition and classification by the analysis of coupling between different physiological signals [41] as well as the El Niño prediction using the analysis of climate network dynamics where the links have been determined by the correlations between the individual nodes given by observational sea surface temperature records [42–45]. Finally, we believe that our solutions could appear useful in shape comparison problems arising in the analysis and interpretation of microscopic images, mass spectrometry data, and many other applications where the stochastic noise exhibits significant and fluctuating levels that cannot be ignored.

ACKNOWLEDGMENTS

We would like to thank Professor Yuri D. Uljanitski for illuminating discussions as well as two anonymous referees for their helpful comments. We also thank the Ministry of Education and Science of the Russian Federation for the financial support of this work within the Basic State Assignment of St. Petersburg Electrotechnical University.

[1] W. Gibson, *Pattern Recognition* (Putnam, New York, 2003).

[2] S. Singh, M. Singh, C. Apte, and P. Perner, *Pattern Recognition and Image Analysis* (Springer, Berlin, 2005).

[3] J. C. Russ, *Computer-Assisted Microscopy: The Measurement and Analysis of Images* (Springer, Berlin, 2012).

[4] R. Basri, L. Costa, D. Geiger, and D. Jacobs, Vision Res. **38**, 2365 (1998).

[5] M. Germann, T. Latychevskaia, C. Escher, and H.-W. Fink, Phys. Rev. Lett. **104**, 095501 (2010).

[6] K. X. Wan, I. Vidavsky, and M. L. Gross, J. Am. Soc. Mass Spectrom. **13**, 85 (2002).

[7] Z. B. Alfassi, J. Am. Soc. Mass Spectrom. **15**, 385 (2004).

[8] M. Sezgin and B. Sankur, *Proceedings of the 2001 International Conference on Image Processing* (IEEE, Piscataway, 2001), Vol. 3, pp. 764–767.

[9] H. Yang, S. T. Bukkapatnam, and R. Komanduri, Phys. Rev. E **76**, 026214 (2007).

[10] D. Sander and J. Klingelhöfer, J. Neurol. **242**, 313 (1995).

[11] G. Katinas, G. Cornélissen, K. Otsuka, E. Haus, E. Bakken, and F. Halberg, Biomed. Pharmacother. **59**, S141 (2005).

[12] G. Katinas, Y. S. Astakhov, S. Sedov, S. Yashin, S. Chibisov, S. Shastun, S. Y. Astakhov, I. Boldyreva, A. Gromyko, T. Merkuryeva *et al.*, World Heart J. **6**, 261 (2014).

[13] A. Fowler, *Mathematical Geoscience* (Springer, Berlin, 2011).

[14] V. N. Livina, Y. Ashkenazy, A. Bunde, and S. Havlin, *In Extremis* (Springer, Berlin, 2011), pp. 266–284.

[15] V. N. Livina and T. Lenton, Cryosphere **7**, 275 (2013).

[16] F. Yang, A. Kumar, M. E. Schlesinger, and W. Wang, J. Clim. **16**, 2419 (2003).

[17] B. Huang and W. Kinsner, *Proceedings of the IEEE CCECE 2002 Canadian Conference on Electrical and Computer Engineering* (IEEE, Piscataway, 2002), Vol. 2, pp. 1105–1110.

[18] V. Tuzcu and S. Nas, *Proceedings of the 2005 IEEE International Conference on Systems, Man and Cybernetics* (IEEE, Piscataway, 2005), Vol. 1, pp. 182–186.

[19] J. Zhang and M. Small, Phys. Rev. Lett. **96**, 238701 (2006).

[20] J. Zhang, J. Sun, X. Luo, K. Zhang, T. Nakamura, and M. Small, Physica D **237**, 2856 (2008).

[21] O. Sayadi and M. Shamsollahi, Physiol. Meas. **30**, 335 (2009).

[22] J. Zhang, K. Zhang, J. Feng, and M. Small, PLoS Comput. Biol. **6**, e1001033 (2010).

[23] B. Kosko and S. Mitaim, Phys. Rev. E **64**, 051110 (2001).

[24] K. Kitajo, D. Nozaki, L. M. Ward, and Y. Yamamoto, Phys. Rev. Lett. **90**, 218103 (2003).

[25] A. Sokolov, I. S. Aranson, J. O. Kessler, and R. E. Goldstein, Phys. Rev. Lett. **98**, 158102 (2007).

[26] H. V. Nguyen and L. Bai, in *Computer Vision—ACCV 2010*, edited by R. Kimmel, R. Klette, and A. Sugimoto, Lecture Notes in Computer Science Vol. 6493 (Springer, Berlin, 2011), pp. 709–720.

[27] A. Karnik, S. Goswami, and R. Guha, *Proceedings of the First Asia International Conference on Modelling & Simulation* (IEEE, Piscataway, 2007), pp. 165–170.

[28] P. de Chazal and R. B. Reilly, *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing* (IEEE, Piscataway, 2003), Vol. 2, pp. II–269.

[29] R. M. Rangayyan, *Biomedical Signal Analysis: A Case-Study Approach* (IEEE/Wiley-Interscience, New York, 2002).

[30] M. Blanco-Velasco, B. Weng, and K. E. Barner, Comput. Biol. Med. **38**, 1 (2008).

[31] D. Zhang, *Proceedings of the 27th Annual International Conference on Engineering in Medicine and Biology Society* (IEEE, Piscataway, 2005), pp. 1212–1215.

[32] H. Cramer, *Mathematical Methods of Statistics* (Princeton University Press, Princeton, 1999).

[33] To simplify the notation, in the following examples we always denote the linear approximation r_1 as simply r wherever it is evident from the context that the approximate expression is used.

[34] J. E. Kolassa, *Series Approximation Methods in Statistics* (Springer, New York, 2006).

[35] M. I. Bogachev, O. V. Mamontov, A. O. Konradi, Y. D. Uljanitski, J. W. Kantelhardt, and E. V. Schlyakhto, Physiol. Meas. **30**, 631 (2009).

[36] A. Bunde, S. Havlin, E. Koscielny-Bunde, and H.-J. Schellnhuber, Physica A (Amsterdam) **302**, 255 (2001).

[37] E. Koscielny-Bunde, J. W. Kantelhardt, P. Braun, A. Bunde, and S. Havlin, J. Hydrol. **322**, 120 (2006).

[38] M. I. Bogachev, J. F. Eichner, and A. Bunde, Pure Appl. Geophys. **165**, 1195 (2008).

[39] A. Bunde, M. I. Bogachev, and S. Lennartz, in *Complexity and Extreme Events in Geoscience*, edited by S. Sharma, A. Bunde, D. Baker, and V. Dimri, Geophysical Monograph Series (AGU, Washington, DC, 2012), pp. 139–152.

[40] A. A. Tsonis and P. J. Roebber, Physica A (Amsterdam) **333**, 497 (2004).

[41] A. Bashan, R. P. Bartsch, J. W. Kantelhardt, S. Havlin, and P. C. Ivanov, Nat. Commun. **3**, 702 (2012).

[42] K. Yamasaki, A. Gozolchiani, and S. Havlin, Phys. Rev. Lett. **100**, 228501 (2008).

[43] A. Gozolchiani, S. Havlin, and K. Yamasaki, Phys. Rev. Lett. **107**, 148501 (2011).

[44] J. Ludescher, A. Gozolchiani, M. I. Bogachev, A. Bunde, S. Havlin, and H. J. Schellnhuber, Proc. Natl. Acad. Sci. USA **110**, 11742 (2013).

[45] J. Ludescher, A. Gozolchiani, M. I. Bogachev, A. Bunde, S. Havlin, and H. J. Schellnhuber, Proc. Natl. Acad. Sci. USA **111**, 2064 (2014).