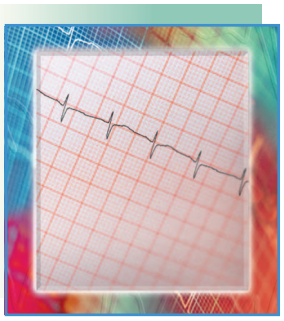

Approximate Entropy for All Signals

Is the Recommended Threshold Value r Appropriate?

BY KI H. CHON,
CHRISTOPHER G. SCULLY,
AND SHENG LU

Calculation of approximate entropy (ApEn) requires a priori determination of two unknown parameters, m and r. While the recommended values of r, in the range of 0.1–0.2 times the standard deviation of the signal, have been shown to be applicable for a wide variety of signals, in certain cases, r values within this prescribed range can lead to an incorrect assessment of the complexity of a given signal. To circumvent this limitation, we recently advocated finding the maximum ApEn value by assessing all values of r from 0 to 1 and found that the maximum ApEn does not always occur within the prescribed range of r values. Our results indicate that finding the maximum ApEn leads to the correct interpretation of a signal's complexity. One major limitation, however, is that the calculation of all choices of r values is often impractical because of the computational burden. Our new method, based on a heuristic stochastic model, overcomes this computational burden and leads to the automatic selection of the maximum ApEn value for any given signal. On the basis of Monte Carlo simulations, we derive general equations that can be used to estimate the maximum ApEn with high accuracy for a given value of m. Application to both synthetic and experimental data confirmed the advantages claimed with the proposed approach.

ApEn is a widely used method to provide a general understanding of the complexity of data [1]. Its popularity stems from the fact that it can be applied to both short- and long-term data recordings, and it is relatively easy to use. Consequently, it has found applications in many disciplines [2], [3].

ApEn determines the conditional probability of similarity between a chosen data segment of a given duration and the next set of segments of the same duration; the higher the probability the smaller the ApEn value, indicating less complexity of the data.

Given a time series with N data points, the calculation of ApEn requires a priori determination of two unknown parameters, m and r. The parameter m determines the length of the sequences to be compared, and its selection can be estimated by calculating the false nearest neighbor [4]. The second parameter, r, is the tolerance threshold for accepting similar patterns between two segments and has been recommended to be within 0.1–0.2 times the standard deviation of the data [1]. This recommendation was largely based on its application to relatively slow dynamic signals such as heart rate [2], [3] and hormonal release data [5]. Our recent work suggests that these recommended values are not always appropriate for fast dynamic neural signals [6]. Furthermore, for a Brownian motion time series, with the selection of $r = 0.15$ times the standard deviation, ApEn value can be low as deterministic signals, which erroneously suggest low complexity of the signal.

Recently introduced variants of ApEn methods, sample entropy (SampEn) [7] and multiscale entropy [8], were developed to overcome the self match problem associated with ApEn and to provide a time-scale-dependent ApEn, respectively. However, these two methods also rely on the choices of both m and r. Therefore, these alternate methods are not immune to the sensitivity of the choice of r.

To this end, our recent work has provided the valuable insight that the most appropriate threshold value is the one that provides the maximum ApEn value [6]. In this study, computer simulation examples consisting of various signals with different complexity were compared. It was found that neither ApEn nor SampEn methods was accurate in measuring signals' complexity when the recommended values (e.g., $m = 2$ and $r =$ 0.1–0.2 times the standard deviation of the signal) were strictly adhered to. However, when we selected the maximum ApEn value as determined by considering many different r values, we were able to correctly discern a signal's complexity for both synthetic and experimental data. However, this requires that many different choices of r values need to be considered. This is a very cumbersome and time-consuming process. Thus, the primary goal of the present work is to illustrate our recently developed method [9] that can automatically select the appropriate threshold value r, which corresponds to the maximum ApEn value, without resorting to the calculation of ApEn for each of the threshold values selected in the range of zero and one times the standard deviation.

Methodology

The details on ApEn algorithm are summarized in [1]. The calculation of ApEn requires a priori specification of two unknown parameters: m, the embedding dimension (ED) and r, a tolerance value. The value of m can be estimated using the

Digital Object Identifier 10.1109/MEMB.2009.934629

 0739-5175/09/$26.00©2009IEEE

The proposed approach provides a more accurate estimate of $ApEn_{max}$ than the conventional method since the difference as denoted by δ is smaller.

calculation of false nearest neighbor [4]. However, theoretical and varied clinical applications, especially for slow dynamics [e.g., heart-rate variability (HRV) and growth hormone release], have shown that either $m = 1$ or 2 and r between 0.1 and 0.2 of the standard deviation of the data provide good statistical validity of ApEn.

Automatic Selection of R That Corresponds to the Maximum ApEn Value

Justification for Choosing the Maximum ApEn Value

To illustrate our reasoning for choosing the maximum ApEn value rather than strictly heeding the recommendation that r be between 0.1 and 0.2 times the standard deviation of the signal, we show ApEn values as a function of r for three different time series with decreasing complexity: white noise (WN), cross chirp, and sinusoidal signals (Figure 1). We note that for all three signals, increasing r at first results in a concomitant increase in ApEn with the maximum ApEn (denoted as $ApEn_{max}$) found at different r values for all three signals. Thereafter, ApEn decreases with increasing r. If we choose $ApEn_{max}$ for all three signals, there is no ambiguity as to which signal is more complex. However, if we were to choose $ApEn_{max}$ based on the recommended r value being between 0.1 and 0.2 times the standard deviation of the signal, the cross-chirp signal (CCS) has a higher ApEn value than the WN signal. Certainly, this is a misleading result. Therefore, this simple example illustrates the pitfalls of the recommended r value selection process, and as an alternative, the most appropriate threshold value, r, can be simply determined by selecting the true $ApEn_{max}$ value. Further, examples demonstrating the appropriateness of selecting $ApEn_{max}$ value, instead of the recommended r value in the range of 0.1–0.2 times the standard deviation of the signal, are provided in our recent study [6]. A recent work using our approach has also confirmed that the threshold value r is critical in human HRV studies [10]. For example, it was found that a selection of $r = 0.25$ resulted in 12% decrease of ApEn value, whereas $r = 0.1$ resulted in 9% increase as subjects changed their body positions from supine to upright. Further, it was found that $ApEn_{max}$ value estimated in human HRV data were consistent within the recommended r values of 0.1–0.2 times the signal standard deviation.

The significance of $ApEn_{max}$ is that it denotes the largest information difference between data length m and $m + 1$ for any given r; thereby, it signifies the maximum complexity. We advocate the use of $ApEn_{max}$ since it is less arbitrary than selecting the recommended r value between 0.1 and 0.2 of the standard deviation of the signal. Furthermore, as shown in Figure 1, even within these suggested r values, there are wide variations in the ApEn values for all three signals, and the results can lead to the incorrect interpretation of complexities between these three signals. However, using $ApEn_{max}$, we obtain the correct information complexities for all three signals. Henceforth, we denote the r value to obtain the $ApEn_{max}$ as the r_{max}.

Automatic Selection of r_{max} Value

To automatically select $ApEn_{max}$ without resorting to the calculation of every possible r value, our method is based on a theory about the behavior of the threshold value. We will show in the subsequent paragraph that the r_{max} value is dependent on the data record length and the square root of the ratio between short- and long-term variability of the signal. To exploit these relationships, the theory begins with a model of a bounded random process (BRP). Most biological signals exhibit both short- and long-term behaviors that can have wide ranges of complexity. The model of the BRP is defined by the following equation:

$$y(i) - y(i-1) = e(i),$$
$$\beta^- < y(i) < \beta^+, \qquad (1)$$

where the time series, $y(i)$ in (1), is the integrated WN signal. Note that the time series $y(i)$ differs from Brownian noise because of the boundaries β defined in the second line of (1). The top expression of (1) describes the short-term variability, in which differences between successive points are assumed to be random processes with the resultant time series, $e(i)$, having zero mean. The standard deviation of $e(i)$ is denoted as sd_1. We use sd_2 to denote the standard deviation of $y(i)$ in (1), which can be thought of as long-term variability of the signal.

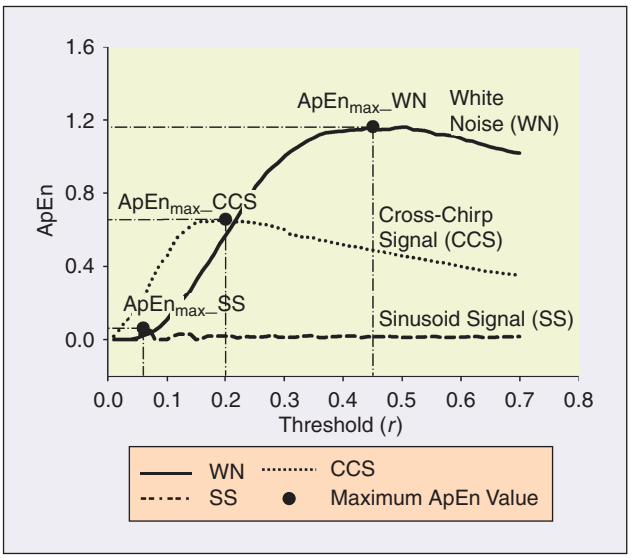

Fig. 1. ApEn values of WN, CCS, and SS with various thresholds (r) are shown.

To simulate wide ranges of complexity, we used Monte Carlo simulations to generate 100 realizations of integrated independent and identically distributed Gaussian WN signals ($e(i) \sim N(0,1)$), with each realization having different bound [$\beta^{\pm}$ is randomly selected from (± 2 to ± 20)]. Thus, 100 realizations resulted in 100 different bounds. For each time series with different data lengths, starting from 200 to 1,000 at an increment of 100, ApEn values corresponding to threshold values ranging from 0.01 to 1, incrementing by 0.01, were computed for different ED values. Only the threshold value that provides the $\mathrm{ApEn}_{\max}$ for a particular ED was selected based on examining all ApEn values resulting from the set of threshold values.

A plot of the optimal r as a function of $\mathrm{sd}_1/\mathrm{sd}_2$ and data length, for EDs two and three, is shown in Figure 2. Figure 2(b) and (d) shows two-dimensional (2-D) illustration of the Figure 2(a) and (c) for data lengths of 200, 600, and 1,000 for the corresponding dimension. While the ApEn is based on choosing r between 0.1 and 0.2 of the standard deviation of the data, we emphasize the use of r as a function of the short-term variability, sd_1, normalized by the long-term variability, sd_2, for our method. We note a quasilinear relationship for both EDs. Thus, either a general linear or nonlinear equation can be derived for each ED. The general equations derived based on fitting multiple nonlinear least squares on the curves shown in Figure 2 are provided below for each ED:

$$
\begin{aligned}
m = 2 &: \hat{r}_{\max} = \left(-0.036 + 0.26\sqrt{\mathrm{sd}_1/\mathrm{sd}_2}\right)/\sqrt[4]{N/1{,}000},\\
m = 3 &: \hat{r}_{\max} = \left(-0.08 + 0.46\sqrt{\mathrm{sd}_1/\mathrm{sd}_2}\right)/\sqrt[4]{N/1{,}000},\\
m = 4 &: \hat{r}_{\max} = \left(-0.12 + 0.62\sqrt{\mathrm{sd}_1/\mathrm{sd}_2}\right)/\sqrt[4]{N/1{,}000},\\
m = 5 &: \hat{r}_{\max} = \left(-0.16 + 0.78\sqrt{\mathrm{sd}_1/\mathrm{sd}_2}\right)/\sqrt[4]{N/1{,}000},\\
\mathrm{m} = 6 &: \hat{r}_{\max} = \left(-0.19 + 0.91\sqrt{\mathrm{sd}_1/\mathrm{sd}_2}\right)/\sqrt[4]{N/1{,}000},\\
m = 7 &: \hat{r}_{\max} = \left(-0.2 + 1.0\sqrt{\mathrm{sd}_1/\mathrm{sd}_2}\right)/\sqrt[4]{N/1{,}000}. \qquad (2)
\end{aligned}
$$

The aforementioned equations differ slightly from those of our previous study [9] largely because of the differences in the random number generators. In addition, we now include new general equations for ED larger than four. We did not consider EDs higher than seven because they are rarely used in practice, as most ApEn users heed to the recommended ED value of two. Higher ED values (>7) can be estimated using the approach we have outlined for interested readers.

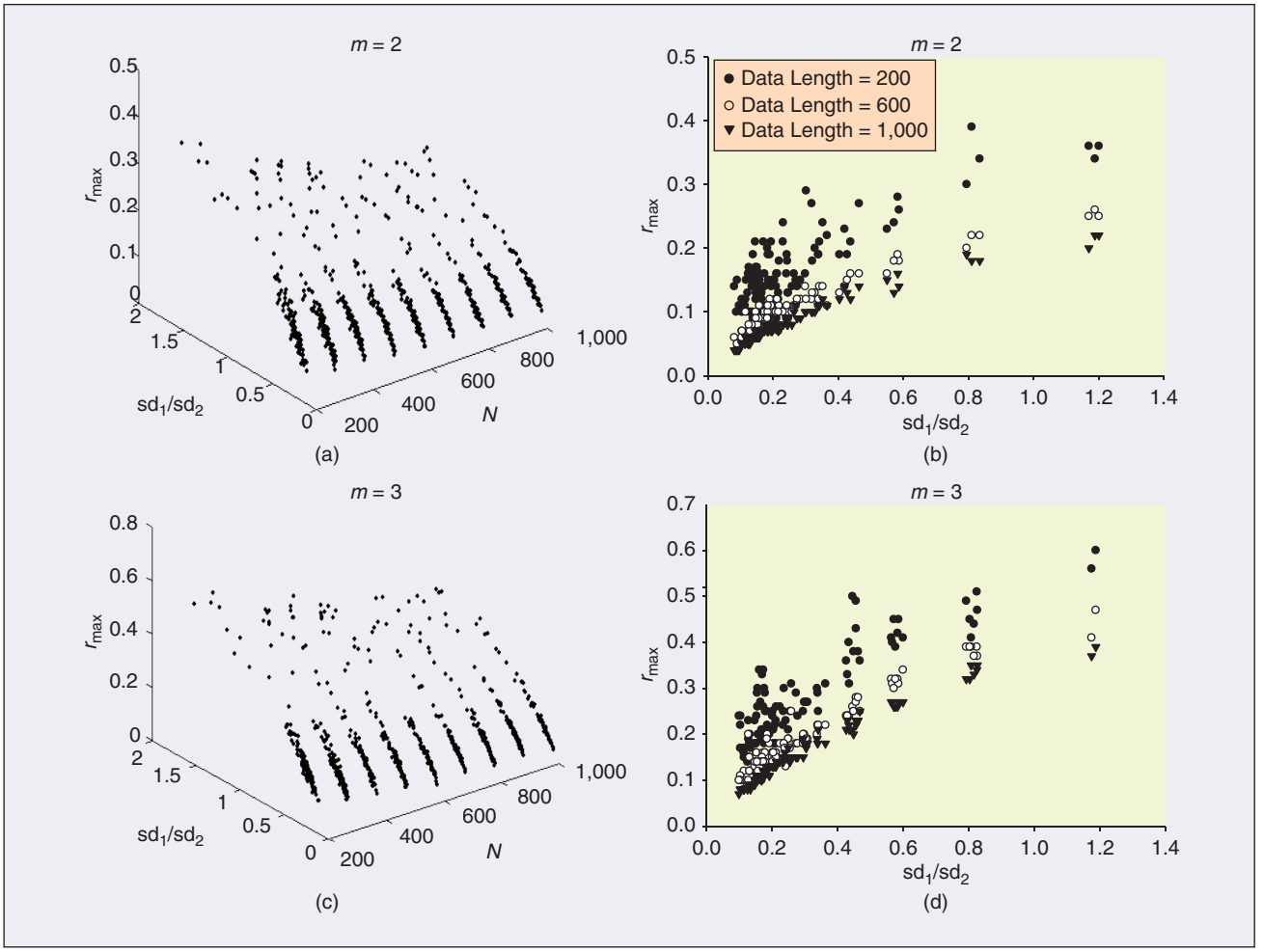

Fig. 2. A Monte-Carlo simulation plot of r_{max} as a function of $\mathrm{sd}_1/\mathrm{sd}_2$ and data length for two different EDs are shown. (b) and (d): 2-D illustrations of (a) and (c) for data lengths of 200, 600, and 1,000 for the corresponding dimension shown. The time series ($N = 1{,}000$) is generated by using the BRP model. r_{max} and $\mathrm{sd}_1/\mathrm{sd}_2$ are calculated for each subset of the time series ($N = 200, 300, \ldots 1{,}000$).

For the aforementioned equations, the estimated r_{max} value approaches a value of zero as N increases to infinity. This should not be of concern since ApEn is not usually calculated for data lengths larger than a few thousand points. An example result, the plot of the optimal r as a function of sd_1/sd_2 for EDs two and three for the data length of 1,000 points, is shown as closed circles and closed inverted triangles in Figure 3. Using (2) and (3), we obtain estimated r_{max} values as a function of sd_1/sd_2, and they are shown as open circles and open inverted triangles in Figure 3. In general, we observe excellent agreement between the actual and estimated r_{max} values although the accuracy degrades a little as the ratio of sd_1/sd_2 increases. Note that the difference between the true and estimated r_{max} does not translate into a large discrepancy between the true and estimated $ApEn_{max}$ values; supporting evidence will be provided in the next section.

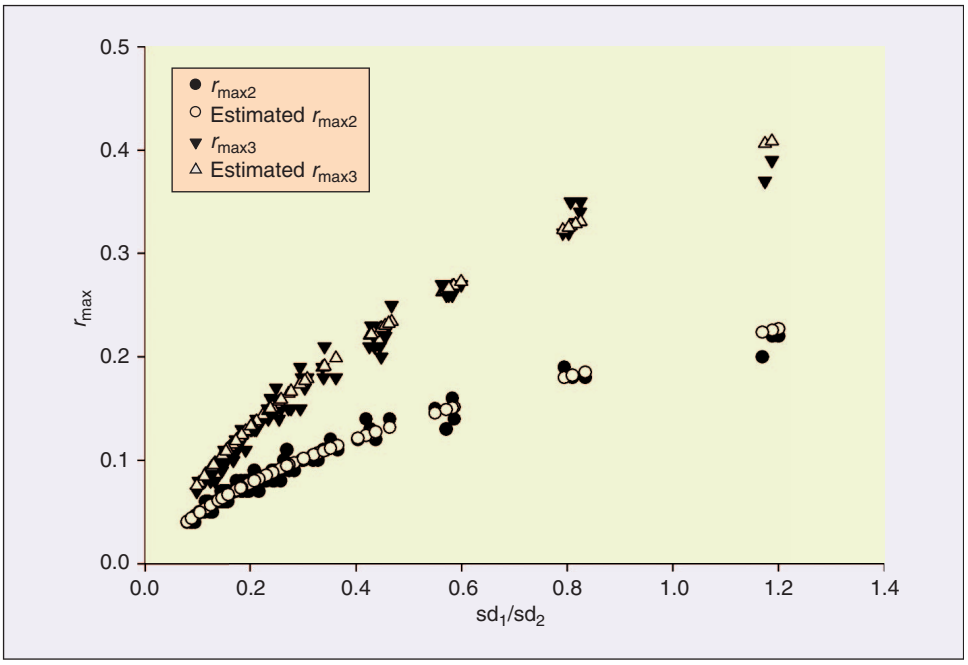

Fig. 3. The plot of the r_{max} as a function of sd_1/sd_2 for two different EDs for the data length of 1,000 points is shown as closed circles ($m = 2$) and closed inverted triangles ($m = 3$). Open inverted triangles ($m = 3$) and open circles ($m = 2$) are estimated values by using (2) and (3).

For experimental data, the ED is estimated first, followed by calculation of sd_1 [standard deviation of $e(i)$ in (1)] and sd_2 [standard deviation of $y(i)$] from which the optimal threshold value is determined from the equations provided for a particular ED.

The solid circle points in Figure 1 represent the actual $ApEn_{max}$ values. For all three signals, m was set to three. The estimated $ApEn_{max}$ values were 1.157 for WN (true $ApEn_{max} = 1.163$), 0.637 for CCS (true $ApEn_{max} = 0.659$), and 0.015 for the SS (true$ApEn_{max} = 0.061$).

Results

Synthetic Simulation Example Consisting of WN, Brownian Motion, Henon, and Logistic Map Series

The results shown are largely derived from our recently published study [9]. To demonstrate the efficacy of our approach in automatically determining the r_{max} value, we generated ten independent realizations each for Gaussian WN, Brownian noise, the logistic map, and Henon map time series. The logistic map is described by the following equation:

$$y(n) - 3.6y(n-1) - 3.6y^2(n-1).$$

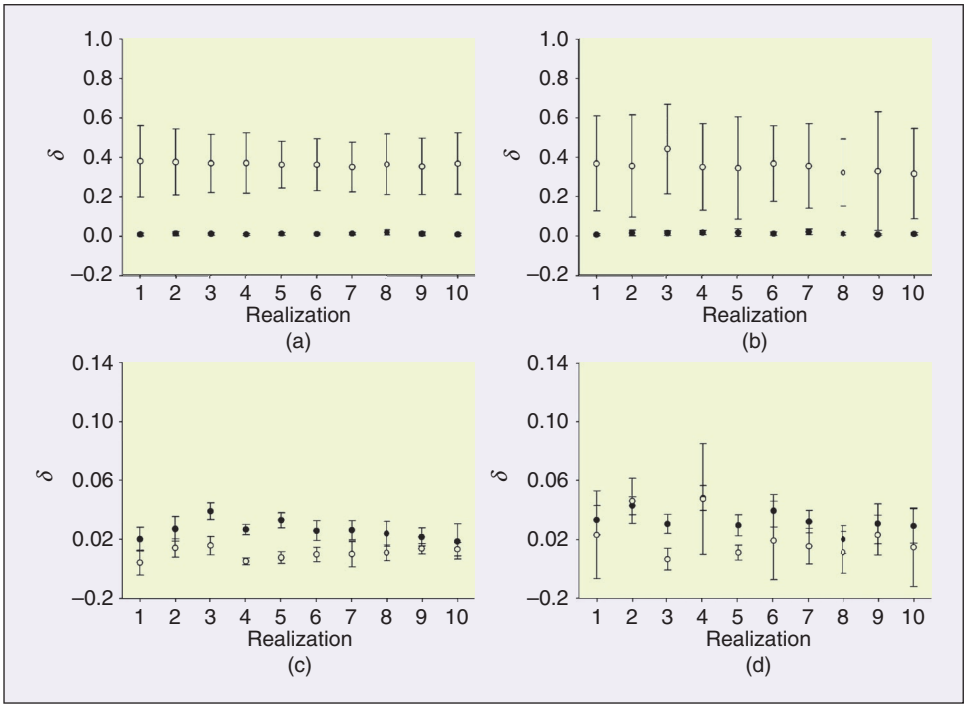

Fig. 4. $|ApEn_{max} - A\hat{p}En_{max}|$ and $|ApEn_{max} - Ap\hat{E}n(2, 0.15)|$ (both are denoted as δ) for Gaussian WN, Brownian motion, logistic map, and Henon map series are shown (ten realization for each type of signal). A solid circle indicates the estimated error by using (2), and an open circle indicates the error estimated by $Ap\hat{E}n(2, 0.15)$.

Most biological signals exhibit both short- and long-term behaviors that can have wide ranges of complexity.

The Henon map is described by:

$$y(n) = 3.168y(n-1) + 0.3y(n-2) - y^2(n-1).$$

Every realization of the four types of signals contained 1,000 data points. From each of these realizations, we generated nine new subrealizations of different data lengths by starting from 200 and incrementing by 100, up to the total data length of 1,000 data points. The purpose of the last step was to examine the variability of the proposed and original ApEn method on different data lengths. For every realization described earlier, the exact $\text{ApEn}_{\max}$ values were determined for every curve as a function of r value, which was successively increased starting from 0.01 times the standard deviation up to one time the standard deviation, at an increment of 0.01. In addition, for every realization, we also determined estimates of $\text{ApEn}_{\max}$ using (2) as well as the conventional estimates of ApEn values based on the arbitrary choice of $m = 2$ and with r set to 0.15 times the standard deviation of the signal. To examine how our proposed approach and the conventional approach compare to the true $\text{ApEn}_{\max}$ values, we calculated the difference between $|\text{ApEn}_{\max} - \text{Ap}\hat{\text{E}}\text{n}_{\max}|$ and $|\text{ApEn}_{\max} - \text{Ap}\hat{\text{E}}\text{n}(2,0.15)|$ for the nine subrealizations to obtain mean and standard deviation values. This procedure was then repeated for each of the remaining nine realizations and their nine subrealizations. The outcome of this comparison is provided in Figure 4. For stochastic signals, the proposed approach provides a more accurate estimate of $\text{ApEn}_{\max}$ than the conventional method since the difference as denoted by δ is smaller. Even for deterministic signals, the proposed method provides a lower magnitude standard deviation around the true $\text{ApEn}_{\max}$ value than the conventional approach. Further, the proposed approach is not affected by the varying data lengths, since both error and standard deviation values are negligible. The conventional approach, however, has higher error and variability than the proposed approach, especially for the stochastic signals.

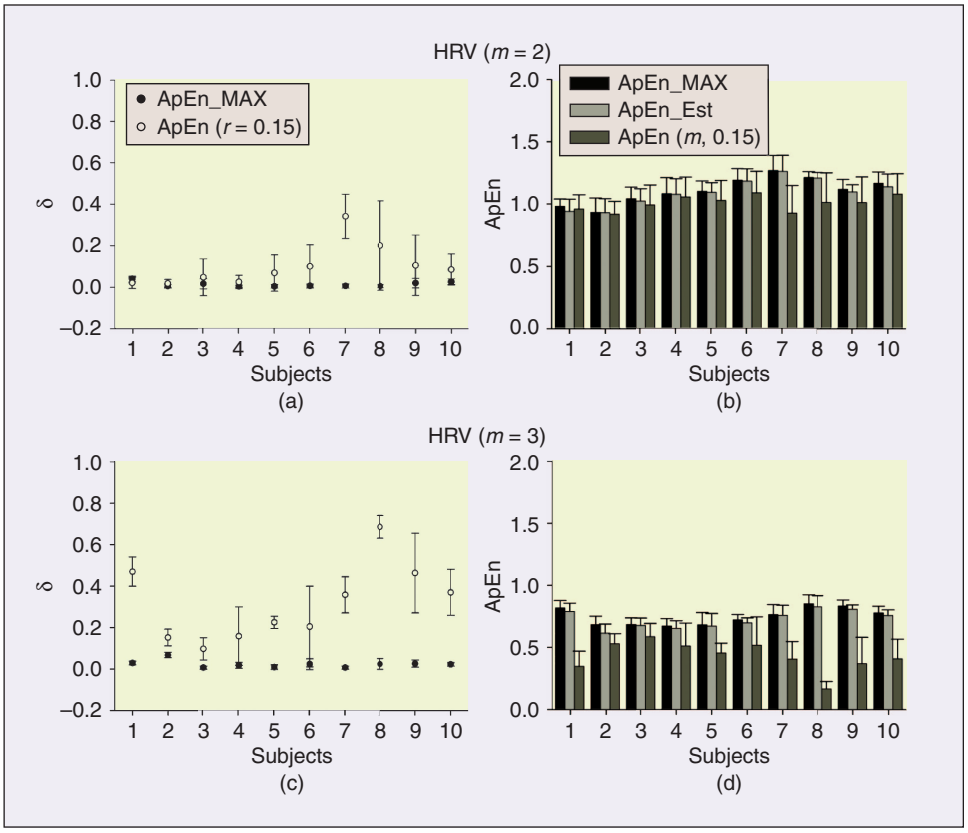

Fig. 5. $|\text{ApEn}_{\max} - \text{Ap}\hat{\text{E}}\text{n}_{\max}|$ and $|\text{ApEn}_{\max} - \text{Ap}\hat{\text{E}}\text{n}(m, 0.15)|$ (both are denoted as δ) for HRV from healthy subjects are shown in (a) and (c) (ten subjects for each type of signal). A solid circle indicates the estimated error by using (2) and (3), and an open circle indicates the error estimated by $\text{Ap}\hat{\text{E}}\text{n}(m, 0.15)$, where $m = 2$ or 3. (b) and (d): Comparison of $\text{Ap}\hat{\text{E}}\text{n}_{\max}$ (black/first column bar) and $\text{Ap}\hat{\text{E}}\text{n}_{\max}(m, 0.15)$ (light gray/second column bar) to the true $\text{ApEn}_{\max}$ (gray/third column bar) for the same ten subjects.

HRV Signals

Biological time series used to demonstrate the efficacy of the automatic selection of $\text{ApEn}_{\max}$ include HRV series from ten healthy subjects (data length $N = 600$). HRV data used in this study consisted of the recordings of surface electrocardiogram (S-ECG). Measurements of S-ECG were sampled at 250 Hz.

For each of these experimental data sets, we performed a similar procedure of generating nine subrealizations

starting from 200, with an increment of 100. We then calculated the difference between $|\mathrm{ApEn}_{\max} - \hat{\mathrm{ApEn}}_{\max}|$ and $|\mathrm{ApEn}_{\max} - \hat{\mathrm{ApEn}}(m, 0.15)|$ for the nine subrealizations. The results from these comparisons are provided in Figure 5(a) and (c) for HRV data. In Figure 5(b) and (d), comparison of $\hat{\mathrm{ApEn}}_{\max}$ and $\hat{\mathrm{ApEn}}(m, 0.15)$ to the true $\mathrm{ApEn}_{\max}$ are provided. For the conventional technique, we used m of either two or three depending on the data set and r of 0.15 times the standard deviation of the signal. For all data sets examined, whether HRV, we noted a negligible difference between the true $\mathrm{ApEn}_{\max}$ and estimated $\mathrm{ApEn}_{\max}$ using our proposed method. Similar to the results obtained with the simulation example above, the conventional ApEn approach resulted in larger errors and greater variability than the proposed approach. Further, even when r was changed to either 0.1 or 0.2 times the standard deviation of the signal to estimate the difference between $|\mathrm{ApEn}_{\max} - \hat{\mathrm{ApEn}}(m, r = 0.1 \text{ or } 0.2)|$, the conventional ApEn resulted in greater variability than the proposed approach. Also illustrated in Figure 5(b) and (d), the estimated $\hat{\mathrm{ApEn}}_{\max}$ values (second column bar of each subject) are much closer to the true $\mathrm{ApEn}_{\max}$ (first column bar of each subject) than $\hat{\mathrm{ApEn}}(m, 0.15)$ (third column bar of each subject) for both $m = 2$ and $m = 3$. The differences between the true $\mathrm{ApEn}_{\max}$ and $\hat{\mathrm{ApEn}}(m, 0.15)$ are much more pronounced especially with $m = 3$ for HRV, whereas they are negligible with $\hat{\mathrm{ApEn}}_{\max}$. It should be noted that, for human data, the closest to the true $\mathrm{ApEn}_{\max}$ is obtained with $m = 2$.

Discussion

To date, determination of ApEn has been made using a recommended r value within the range of 0.1–0.2 times the standard deviation of the signal [1]. This recommendation was largely derived for slow dynamics signals, and the user was left with an arbitrary choice of r value within the range defined earlier [1]. However, as we have shown in this study, ApEn values vary significantly even within the defined range of r values. Furthermore, WN signals have smaller ApEn than chirp signal for some of the recommended r values (see Figure 1). A consequence of this lower ApEn value for WN than Brownian noise is that it leads one to make an incorrect interpretation that the former is less complex than the latter. In an attempt to resolve this inherent deficiency, our recent study suggested that the most appropriate solution is to look for the $\mathrm{ApEn}_{\max}$ value [6]. However, an intractable side effect of finding the $\mathrm{ApEn}_{\max}$ value is the computation of ApEn for every possible r value, which is computationally burdensome.

In this review, we highlight our recently developed method to automatically determine the $\mathrm{ApEn}_{\max}$ without resorting to calculation of ApEn for every possible r value [9]. The method is based on a heuristic model termed the BRP, which is a stochastic model that incorporates characteristics of both short- and long-term variability inherent in the signal. Using Monte Carlo simulations, we derived the general equations for determining the $\mathrm{ApEn}_{\max}$ $m = 2 - 7$. The derived equations also took into account data length dependency of ApEn.

While the BRP model is a stochastic model, we have validated the accuracy of the proposed approach in finding the $\mathrm{ApEn}_{\max}$ values for many varieties of deterministic signals. Figures 4 and 5 illustrate the efficacy of the proposed approach as the difference between the actual $\mathrm{ApEn}_{\max}$, and the estimated $\mathrm{ApEn}_{\max}$ values are negligible for a wide variety of signals (including deterministic signals), and the accuracy remains unaffected by different data lengths. Thus, the burden of automatically finding the $\mathrm{ApEn}_{\max}$ has been averted by (2)–(4). Furthermore, these equations lead to a more accurate representation of the complexity than does the conventional method for most deterministic and stochastic signals.

Ki H. Chon is a professor in the Department of Biomedical Engineering at the State University of New York (SUNY) Stony Brook. His research interests include medical instrumentation, biomedical signal processing, and identification and modeling of physiological systems.

Christopher G. Scully received his B.S. degree in electromechanical engineering from Wentworth Institute of Technology. He is currently a Ph.D. student at SUNY Stony Brook in the Department of Biomedical Engineering, working with Dr. Chon. His research interests include clinical applications of biological signal processing and medical instrumentation.

Sheng Lu received his M.S. and Ph.D. degrees in electrical engineering from the City University of New York, New York, in 2000 and 2002, respectively. He joined Sterlingtech Software Inc. in 2008 as a medical device software consultant. His research interests include cardiovascular signal processing, system identification, modeling of physiological systems, and medical software development.

Address for Correspondence: Ki H. Chon, Department of Biomedical Engineering, SUNY Stony Brook, Stony Brook, New York City, New York, USA. E-mail: ki.chon@sunysb.edu.

References

[1] S. M. Pincus, "Approximate entropy as a measure of system complexity," *Proc. Natl. Acad. Sci. USA*, vol. 88, pp. 2297–301, Mar. 1991.

[2] K. K. Ho, G. B. Moody, C. K. Peng, J. E. Mietus, M. G. Larson, D. Levy, and A. L. Goldberger, "Predicting survival in heart failure case and control subjects by use of fully automated methods for deriving nonlinear and conventional indices of heart rate dynamics," *Circulation*, vol. 96, pp. 842–848, Aug. 1997.

[3] D. T. Kaplan, M. I. Furman, S. M. Pincus, S. M. Ryan, L. A. Lipsitz, and A. L. Goldberger, "Aging and the complexity of cardiovascular dynamics," *Biophys. J.*, vol. 59, pp. 945–949, Apr. 1991.

[4] M. B. Kennel, R. Brown, and H. D. Abarbanel, "Determining embedding dimension for phase-space reconstruction using a geometrical construction," *Phys. Rev. A*, vol. 45, pp. 3403–3411, Mar. 1992.

[5] S. M. Pincus, M. L. Hartman, F. Roelfsema, M. O. Thorner, and J. D. Veldhuis, "Hormone pulsatility discrimination via coarse and short time sampling," *Amer. J. Physiol.*, vol. 277, pp. E948–57, Nov. 1999.

[6] X. Chen, I. C. Solomon, and K. H. Chon, "Parameter selection criteria for approximate entropy and sample entropy with application to neural-respiratory signals," submitted for publication.

[7] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *Amer. J. Physiol. Heart Circ. Physiol.*, vol. 278, pp. H2039–H2049, June 2000.

[8] M. Costa, A. L. Goldberger, and C. K. Peng, "Multiscale entropy analysis of complex physiologic time series," *Phys. Rev. Lett.*, vol. 89, p. 068102, Aug. 2002.

[9] S. Lu, X. Chen, J. K. Kanters, I. C. Solomon, and K. H. Chon, "Automatic selection of the threshold value R for approximate entropy," *IEEE. Trans. Biomed. Eng.*, vol. 55, pp. 1966–1972, Aug. 2008.

[10] P. Castiglioni and M. Di Rienzo, "How the tolerance threshold "R" influences approximate entropy analysis of heart rate variability," in *Computers in Cardiology*, Bologne, Italy, 2008, pp. 561–564.