SEPARATION OF DETERMINISTIC AND STOCHASTIC COMPONENTS FROM TIME SERIES*

Monika Petelczyc

Faculty of Physics, Warsaw University of Technology
Koszykowa 75, 00-662 Warszawa, Poland

Jakub M. Gac

Faculty of Chemical and Process Engineering, Warsaw University of Technology
Waryńskiego 1, 00-645 Warszawa, Poland

(Received March 10, 2014)

In this study, we propose a modification of a method for the extraction of the dynamics of a system from a time series generated by it. We focused on the problem of rare events for which a statistical characterization is difficult because of their small number. We developed a method for the separation of the deterministic and stochastic components of the time series through the computation of probability densities. Our previous method used a constant width of bins in the histograms for determination of the probability densities. Here, we replace them by bins with a constant number of counts. We have tested the method and presented an application to heart rate variability showing advantages of the modified procedure.

DOI:10.5506/APhysPolBSupp.7.395
PACS numbers: 02.50.Ey, 02.50.Fz, 05.10.Gg, 87.10.Ed

1. Introduction

The analysis of experimental data is difficult not only because of non-stationarity, length of the time series, the occurrence of artifacts but also because of the occurrence of rare events. The statistical insignificance of a rare events result in removal or their replacement. It is well known that characteristics of extreme events may be crucial in prediction, the modelling of the processes or in a simple fluctuation analysis of the data [1]. Another problem with real time series is the determination of the dynamics of a system which is disturbed by noise. While measurement noise [2] (random

* Presented at the Summer Solstice 2013 International Conference on Discrete Models of Complex Systems, Warszawa, Poland, June 27–29, 2013.

variable added to each result of measurement) is relatively easy to remove, dynamical noise [3] (which interacts with the system) is difficult to recognize and separate. In this paper, we introduce a method for the extraction of the deterministic and stochastic components from a discrete time series. This is our second, expanded study which contains a modified extraction procedure taking into account rare events.

2. Description of the method

The method of the extraction of the system dynamics from noisy time series has been developed in our previous work [4]. Here, we modify the method for the separation of the deterministic and stochastic components to enable the study of extreme events in the data. Usually, the extreme values in time series are excluded from the analysis because of their low statistical significance. Our work is an attempt to include the rare events in the study without filtering them or their replacement. We begin with a short description of this method.

Consider a one-dimensional discrete time system with noise. The equation describing the system has the form

$$x_{n+1} = f(x_n) + g(x_n)\,\xi_n\,. \tag{1}$$

To simplify the notation, we rewrite Eq. (1) into

$$x' = f(x) + g(x)\,\xi\,, \tag{2}$$

where $f(x)$ denotes the deterministic part and $g(x)$ stochastic part of the dynamics. In equations (1), (2) ξ denotes the noise term. In the present paper, we assume that the noise in (1) is uncorrelated. However, the generalization of our method for the case of correlated noise is now in progress. In [4], we showed that if ξ has a stationary distribution with the first two moments equal to zero and one, respectively, the functions $f(x)$ and $g(x)$ may be computed from the formulas

$$f(x) = \int_{-\infty}^{+\infty} x' q(x' \mid x)\, dx' \tag{3}$$

and

$$g(x) = \sqrt{\int_{-\infty}^{+\infty} x'^2 q(x' \mid x)\, dx' - f^2(x)}\,, \tag{4}$$

where $q\,(x' \mid x)$ is the conditional probability density. The final forms of $f\,(x)$ and $g\,(x)$ functions depend on the computation of this density, therefore, we present the following procedure for the determination $q\,(x' \mid x)$ from real time series data.

First, we divide the range of the signal into bins. In the basic method [4], we assumed that all the bins have the same width, *i.e.* if the number of bins is equal to N and the minimum and maximum values of the time series are equal to $x_{\min}$ and $x_{\max}$, respectively, the width of every bin is equal to $\delta = \frac{x_{\max} - x_{\min}}{N}$ [4]. Next, we search for all the pairs $\{x_k, x_{k+1}\}$ for which $x_k = x$. This means binning the data and choosing such pairs of points $\{x_k\}$ and $\{x_{k+1}\}$ which fall into the same bins with x and x'. If the number of these pairs is given by N_j, and N_i is the number of the points $\{x_k\} = x$ in the time series, the conditional probability is computed from

$$Q\left(x_{k+1} \in \left(x' - \frac{\delta}{2},\ x' + \frac{\delta}{2}\right) \mid x_k \in \left(x - \frac{\delta}{2},\ x + \frac{\delta}{2}\right)\right) = \frac{N_j}{N_i}\,. \tag{5}$$

The conditional probability density function $q\,(x' \mid x)$ can be now calculated from the formula

$$q\left(x' \mid x\right) \cong \frac{N_j}{N_i \delta}\,. \tag{6}$$

For each bin of width δ, the conditional probability Q is determined and in each the probability density is given by (6). Finally, the integrals in Eqs. (3) and (4) may be computed by means of the trapeze rule. It is obvious that the accuracy of calculation of the functions $f\,(x)$ and $g\,(x)$ depends on the width of bins and on the average number of pairs in every bin. This issue was discussed in details in [4]. In our previous work, we also presented the results of the application of the method to the logistic and tent map as well as to heart rate variability data. In the case of the chaotic maps, we obtained a satisfactory agreement between reconstructed $f\,(x)$, $g\,(x)$ and the original functions used for generation of the time series. However, we observed a weak point of the method which manifests itself in a poor reconstruction of the $f\,(x)$ and $g\,(x)$ functions for extreme values of RR intervals in the real time series. This is the result of the low number of data points (RR intervals) in these outside bins. In that regions the functions $f\,(x)$ and $g\,(x)$ exhibit strong fluctuations which originate exactly from low number of data points. Therefore, the extraction of the noise ξ for both extreme ranges of the argument may not be accurate. To avoid these problems, we suggest the modification of our method. The main idea of this modification is division of time series into uneven bins. In the new division, every bin contains the same number of counts. As a result, the width of every bin is, in general, different. Usually, the bins near $x_{\min}$ and $x_{\max}$ are broader than the others. The density probability in Eq. (6) is now assigned to the middle of the bin.

Finally, the functions $f(x)$ and $g(x)$ are calculated according to Eq. (6), by means of the generalized trapeze rule applied to the nonequal bins.

A specific difficulty may appear in the series in which repetitive values appear, *e.g.* $x_i = y$ for more than one i. In such series, all x_i values get into the one bin and then this bin has too many countings. Moreover, these repetitive values are NOT placed into surrounding bins. A simple redistribution into neighbouring bins will result in decreasing number of elements in the middle bin. To manage this problem, we add to every element of repetitive values a low random variable (white Gaussian noise — the precision about its magnitude will be given later) which eliminates the number of pairs for which $x_i = x_j$.

Having reconstructed the functions $f(x)$ and $g(x)$, it is possible to reconstruct the time series for the noise component ξ and then its distribution. ξ from Eq. (2) is

$$\xi = \frac{x' - f(x)}{g(x)}\,. \tag{7}$$

Moreover, the statistical properties of the noise ξ may be used to estimate the accuracy of the extraction of the dynamics components from the time series, when the original deterministic and stochastic terms are not known. This can be done by comparison of the mean and variance of the extracted noise and the values: zero and one respectively. These values were assumed in Eq. (2), which describes the dynamics of the system.

3. Test of the modified method

We tested our method of denoising with the constant width of bins in the conditional probability density on two time series with different type of noise: Gaussian and Gumbel [4]. Here, we test the method on the logistic map with the Gumbel noise [5] added and assuming a zero mean and the variance equal to one. We used equation

$$x_{n+1} = 2.13\, x_n\,(1 - x_n) + (0.056\, x_n + 0.02)\,\xi_n\,. \tag{8}$$

The return map for the series obtained from this equation is presented in Fig. 1. Here, ξ_n is the Gumbel noise with the probability density function (PDF) given by

$$P(y) = \frac{1}{B}\exp\left(\frac{A-y}{B}\right)\exp\left[-\exp\left(\frac{A-y}{B}\right)\right]\,. \tag{9}$$

The parameters $A, B > 0$ define the mean and the variance. They were selected to obtain the assumed properties of moments of the noise distribution. The results of the reconstruction of the functions $f(x)$ and $g(x)$ are presented in Table I and in Fig. 2.

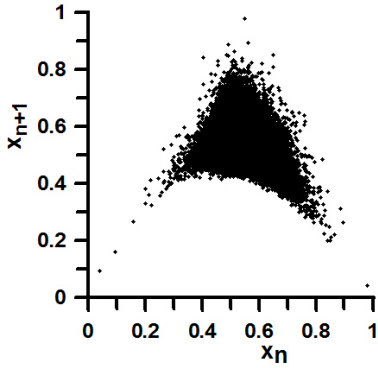

Fig. 1. Return map for the logistic map with Gumbel noise. Time series contains 10^5 points.

TABLE I

The results for $f(x)$ and $g(x)$ for the logistic map with Gumbel noise for a 10^5 points in time series. Computation were done for 100 bins of unequal width.

	$f(x)$	$g(x)$
Original	$-2.13\,x^2 + 2.13\,x$	$0.056\,x + 0.02$
Reconstructed	$-2.08\,x^2 + 2.11\,x - 0.001$	$0.050\,x + 0.030$

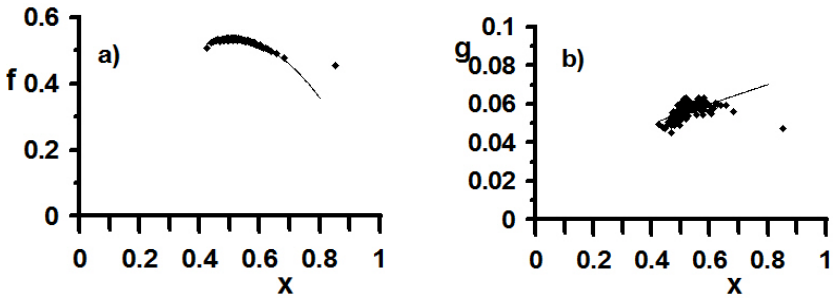

Fig. 2. Functions (a) f and (b) g determined using the modified method of denoising for the logistic map with Gumbel noise. The equations of fitted curve and line are presented in the last row of Table I.

We observed that the current method gives an increased accuracy for the function $f(x)$ in comparison to the previous method [4] with the constant width of bins. The quadratic and linear terms are larger and yield results which are closer to the original terms while the constant term is smaller.

The better agreement of the original and the reconstructed functions $f(x)$ was obtained because here the fitting range was extended in comparison to the previous method. Unfortunately, the accuracy of $g(x)$ is not enhanced but note that the magnitude of the noise part in Eq. (8) is very low. In addition, knowing the noise ξ used in generation of the time series for the logistic map, we compared the reconstructed noise and original one (see the distributions in Fig. 3).

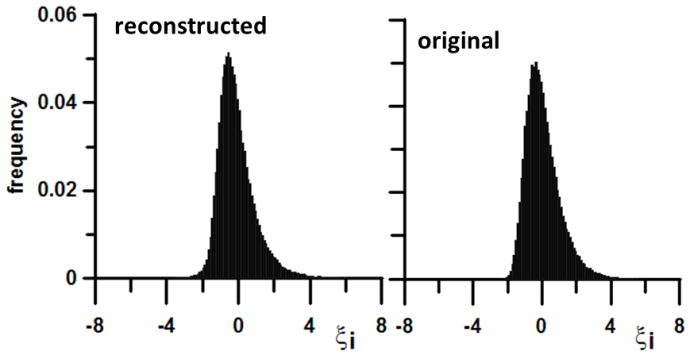

Fig. 3. Distribution for the reconstructed noise ξ determined using discrete method of denoising from logistic map with Gumbel noise and the distribution of the original noise ξ used for generation of time series.

The basic statistical properties of the distributions of ξ are given in Table II. The reconstruction of the noise was done directly from the extracted $f(x)$ and $g(x)$. From the differences in the minima and maxima of ξ, we conclude that the largest deviations result from the outer regions of $f(x)$ and $g(x)$ functions. We decided to interpolate $f(x)$ and $g(x)$ by fitting procedure for noise reconstruction. From our procedure we determine the $f(x)$ and $g(x)$ only for the middle of every bin. As a result of the interpolation

TABLE II

The skewness, kurtosis, minima and maxima for distribution of ξ obtained from the reconstruction and the original one used in the generation of the time series.

Parameter	Reconstructed	Original
Skewness	1.28	1.10
Kurtosis	−20.38	−20.57
Minimum	−3.48	−2.48
Maximum	8.95	8.92
SD	1.03	1.00

(given by polynomial fitting in windows) we obtain continuous $f(x)$ and $g(x)$ functions for whole range of argument. Then, we compute the values of ξ for every point x_i using formula (7) with $x = x_i$ and $x' = x_{i+1}$.

4. Application to heart rate variability

In this section, we present application of the method to a real time series. We focused on heart rate variability because usually, for healthy persons, the extreme events are rare. On the other hand, there are many heart diseases in which arrhythmias occur and result in pairs of a short RR interval followed by an interval about twice the local average — *i.e.* extreme events. We expect that the dynamics obtained using the method described here, from recordings with different numbers of arrhythmias, will have different properties and will be characterized by a particular form of the functions $f(x)$, $g(x)$ and by the noise ξ.

We compared two signals of heart rate variability from male patients with aortic valve stenosis, one has an ejection fraction 71% (time series are named by ST acronym) and the second a very low one — 40% (time series are named by LST acronym). Patients were age 25 and 23 y, respectively. We analysed 6 hour night-time data sets recorded between 9 p.m. and 6 a.m. The heart rate variability time series — the RR intervals [6] — were extracted from a 24 h Holter ECG recording using the Del Mar Reynolds system (Spacelabs) at the Institute of Cardiology (Warsaw, Poland). The data were checked manually by a cardiologist: normal beats were detected, artifacts were deleted by hand. For our purpose, it is crucial that no arrhythmia filtering was applied as the occurrence of arrhythmias is a factor contributing to the level of noise. The series of RR intervals contain many repetitive values because of the relatively low sampling frequency — here 128 Hz — and the fact that heart rhythm fluctuates around a certain level during the night. We added small Gausian distributed random variable to maintain the constant number of points in each bin. We proposed the standard deviation of that noise smaller than the sampling resolution. If the data are sampled at 128 Hz then the sampling resolution is 8 ms, so the standard deviation of the random variable should be assumed at least $1/6^{\text{th}}$ of 8 ms. We decided to use this fraction because in the range of 6σ there are 99.7% of Gaussian distributed variables, and yields a smooth redistribution of the RR intervals.

The comparison of the functions $f(x)$ and $g(x)$ obtained from the earlier version of the method and from the modified version described here is shown in Fig. 4. Many outlier points (RR $<$ 900 ms) in $f(x)$ obtained from previous version of the method (Fig. 4 (b)) are integrated into only a few points in the $f(x)$ function reconstructed using new method (Fig. 4 (d) marked by

arrows). The $g(x)$ from Fig. 4 (c) fluctuates for short and long extremes of argument range. Therefore, computation of the noise ξ is impossible, because we can do fitting for the $g(x)$ only for a narrow range of the argument. Our modified method gives better results for $g(x)$ (see Fig. 4 (e)). Although it is difficult to fit a single function for $g(x)$, we may now apply piecewise fitting. Simultaneously, we also do not remove the few outliers from further analysis as they are critical for occurrence of rare events.

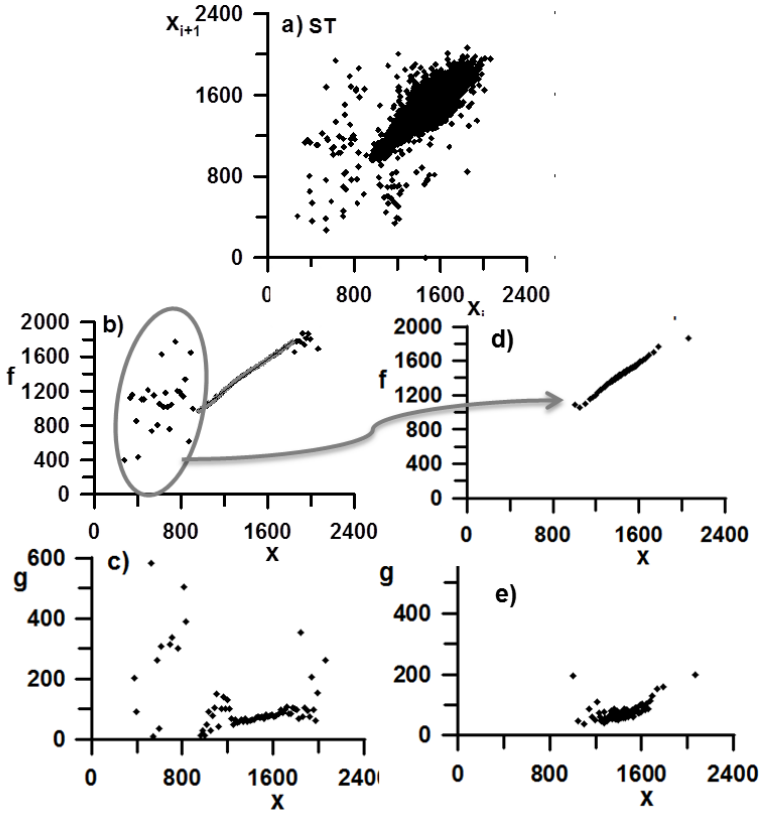

Fig. 4. (a) Return map for the heart rate variability for the patient ST. Comparison of the dynamics for the deterministic and the stochastic terms reconstructed using the method with the constant width of the bins (b), (c) and with the constant number of points in each bin (d), (e).

The dominance of the modified method is especially seen for heart rate variability from the patient LST (Fig. 5 (a)), for whom we observe a few clusters (RR $<$ 800 ms and RR $>$ 1500 ms) around the preserved "comet" of points [7]. The comet shape is typical for heart rate variability of healthy

people. Comet shape is seen as a group of points in the return map. This group of points is concentrated in narrow space for short RR intervals and then extends for longer RR intervals. The clouds of points outside the comet are related to large accelerations or decelerations of heart rate and may be signs of pathology. For the corresponding range of the argument our previous method does not work well. $f\,(x)$ fluctuates strongly and that results in large variations of $g\,(x)$ (Fig. 5 (c), (e)).

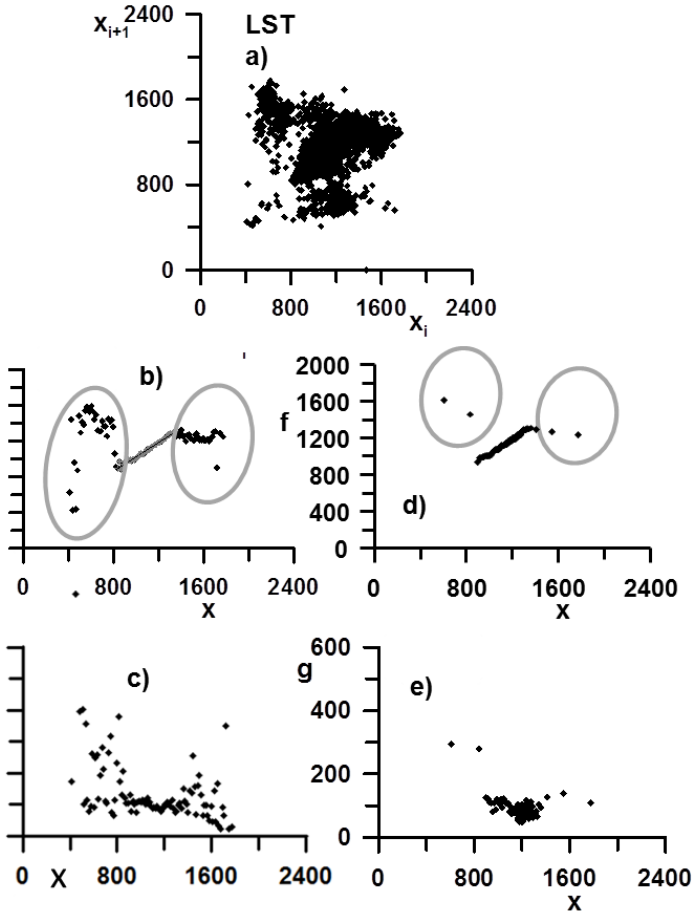

Fig. 5. (a) Return map for heart rate variability for the patient LST. Comparison of the dynamics for deterministic and stochastic terms reconstructed using the method with constant bin width (b), (c) and with the constant number of counts in the bins (d), (e).

Finally, we compared the results of the reconstruction of the dynamics described by the function $f(x)$ and the obtained statistical parameters of ξ distribution for the two time series: ST and LST. The results are presented in Table III. Note, that the reconstruction of noise was done after interpolation (piecewise fitting) of both functions $f(x)$ and $g(x)$. Differences between the properties of the two signals were obtained, especially in skewness and kurtosis of distribution of ξ. We observed a similarity in $f(x)$ for the two cases of stenosis patients. However, the statistical parameters of the ξ distribution differ significantly. We expect that these parameters may indicate an advancement of the disease. This hypothesis should be verified by the analysis of numerous HRV series of various stenosis patients.

TABLE III

Comparison of $f(x)$ and the statistical parameters for the extracted noise of the heart rate variability for ST and LST. The results were obtained from the method with constant number of counts per bin. For the computation of the functions ξ $f(x)$ and $g(x)$ were interpolated.

Case	$f(x)$	Skewness of ξ distribution	Curtosiss of ξ distribution
ST	$0.85\,x + 208.7$	-0.208	-1.025
LST	$0.65\,x + 425.1$	0.968	7.827

5. Conclusions

In this paper, we presented a method of extraction of the deterministic and stochastic components from discrete time series. This method differs from our earlier method described in detail in [4]. The difference is given by the new probability density construction. Previously, we used histograms with equal bins width. In the present method, the bins do not have an equal width. They contain an equal number of counts instead. In this study, we showed that this modification allows us to avoid the difficulties of reconstruction of deterministic and stochastic part near both ends of X range. Indeed, in the original method, the marginal bins contain very few data points that results in fluctuations of computed functions f and g. We observed that for the modified method these fluctuations do not appear or have much smaller magnitude. The application of both methods to the artificial time series (in our case: the logistic map) showed that both the methods give quite similar results for the forms of the deterministic and stochastic components. The differences obtained from the analysis of real data (HRV signals) are more significant. Not only the forms of the functions $f(x)$ and $g(x)$ are different but also is the distribution of reconstructed

noise ξ. However, both methods have been developed with the assumption that the time series are one-dimensional and the noise ξ is independent of x. These limitations may not be appropriate for experimental data, therefore, a generalization of the methods should be done. This generalization will be the subject of our forthcoming paper.

The authors would like to thank R. Baranowski from the Institute of Cardiology in Warszawa for allowing us to use the medical data. We thank Prof. J.J. Zebrowski for discussions and useful remarks. This work was supported by the Polish National Center for Scientific Research grant No. 2011/03/B/ST2/03695 — "Fluctuations and Nonlinear Phenomena in the Human Cardiovascular System — New Methods of Analysis and Modeling".

REFERENCES

[1] S. Albeverio, J. Jentsch, H. Kantz, Extreme Events in Nature and Society, Center for Frontiers Sciences, 2010.

[2] P. Grassberger *et al.*, *Chaos* **3**, 127 (1993).

[3] K. Urbanowicz, J.A. Hołyst, *Phys. Rev.* **E67**, 046218 (2003).

[4] M. Petelczyc, J.M. Gac, J.J. Żebrowski, *Phys. Rev.* **E86**, 011114 (2012).

[5] E. Bertin, *Phys. Rev. Lett.* **95**, 170601 (2005).

[6] G.B. Moody, RR Intervals, Heart Rate, and HRV Howto, Physionet, `http://www.physionet.org/tutorials/hrv/`

[7] M.A. Woo *et al.*, *Am. Heart J.* **123**, 704 (1992).