Regeneration of stochastic processes: an inverse method

F. Ghasemi[1], J. Peinke[2], M. Sahimi[3], and M.R. Rahimi Tabar[1,4,a]

[1] Department of Physics, Sharif University of Technology, PO Box 11365-9161, Tehran 11365, Iran
[2] Carl von Ossietzky University, Institute of Physics, 26111 Oldenburg, Germany
[3] Department of Chemical Engineering, University of Southern California, Los Angeles, California 90089-1211, USA
[4] CNRS UMR 6529, Observatoire de la Côte d'Azur, BP 4229, 06304 Nice Cedex 4, France

Abstract. We propose a novel inverse method that utilizes a set of data to construct a simple equation that governs the stochastic process for which the data have been measured, hence enabling us to reconstruct the stochastic process. As an example, we analyze the stochasticity in the beat-to-beat fluctuations in the heart rates of healthy subjects as well as those with congestive heart failure. The inverse method provides a novel technique for distinguishing the two classes of subjects in terms of a drift and a diffusion coefficients which behave completely differently for the two classes of subjects, hence potentially providing a novel diagnostic tool for distinguishing healthy subjects from those with congestive heart failure, even at the early stages of the disease development.

PACS. 05.10.Gg Classical ensemble theory – 05.40.-a Fluctuation phenomena, random processes, noise, and Brownian motion – 05.45.Tp Time series analysis

1 Introduction

Many natural or man-made phenomena, as well as the morphology of numerous physical systems, are characterized by a degree of stochasticity. Turbulent flows, fluctuations in the stocks prices, seismic recordings, the internet traffic, pressure fluctuations in chemical reactors, and the surface roughness of many materials and rock [1,2] are but a few examples of such phenomena and systems. A long standing problem has been the development of an effective reconstruction method for such phenomena. That is, given a set of data for certain characteristics of a phenomenon, one would like to develop an effective equation that can reproduce the data with an accuracy comparable to the measured data. If such a method can be developed, one may utilize it to, (1) reconstruct the original process with similar statistical properties, and (2) understand the nature and properties of the stochastic process.

In this paper we use a novel method to address this general problem. The proposed method utilizes a set of data for a phenomenon which contains a degree of stochasticity and constructs a simple equation that governs the phenomenon. As we show below, in addition to being highly accurate, the method is quite general; it is capable of providing a rational explanation for complex features of the phenomenon; it requires no scaling feature, and it enables us to accomplish the tasks listed above. As an example, we apply the method to analyze cardiac interbeat intervals which normally fluctuate in a complex

manner. We show that the application of the method to the analysis of interbeat fluctuations in the heart rates may potentially lead to a novel method for distinguishing healthy subjects from those with congestive heart failure (CHF).

2 The method

We begin by describing the steps that lead to the development of a stochastic equation, based on the (stochastic) data set, which is then utilized to reconstruct the original data, as well as an equation that describes the phenomenon.

(1) As the first step we check whether the data follow a Markov chain and, if so, estimate the Markov time (length) scale t_M. As is well-known, a given process with a degree of randomness or stochasticity may have a finite or an infinite Markov time (length) scale. The Markov time (length) scale is the minimum time interval over which the data can be considered as a Markov process [3–6]. To determine the Markov scale t_M, we note that a complete characterization of the statistical properties of stochastic fluctuations of a quantity x in terms of a parameter t requires the evaluation of the joint probability distribution function (PDF) $P_n(x_1, t_1; \cdots; x_n, t_n)$ for an arbitrary n, the number of the data points. If the phenomenon is a Markov process, an important simplification can be made, as the n-point joint PDF, P_n, is generated by the product of the conditional probabilities $p(x_{i+1}, t_{i+1}|x_i, t_i)$, for

[a] e-mail: rahimitabar@iust.ac.ir

$i = 1, \cdots, n - 1$. A necessary condition for a stochastic phenomenon to be a Markov process is that the Chapman-Kolmogorov (CK) equation [7],

$$p(x_2, t_2|x_1, t_1) = \int \mathrm{d}(x_3) \ p(x_2, t_2|x_3, t_3) \ p(x_3, t_3|x_1, t_1), \tag{1}$$

should hold for any value of t_3 in the interval $t_2 < t_3 < t_1$. One should check the validity of the CK equation for different x_1 by comparing the directly-evaluated conditional probability distributions $p(x_2, t_2|x_1, t_1)$ with the ones calculated according to right side of equation (1). The simplest way to determine t_M for stationary or homogeneous data is the numerical calculation of the quantity, $S = |p(x_2, t_2|x_1, t_1) - \int \mathrm{d}x_3 p(x_2, t_2|x_3, t_3) p(x_3, t_3|x_1, t_1)|$, for given x_1 and x_2, in terms of, for example, $t_3 - t_1$ and considering the possible errors in estimating S. Then, $t_M = t_3 - t_1$ for that value of $t_3 - t_1$ for which S vanishes or is nearly zero (achieves a minimum).

(2) Deriving an effective stochastic equation that describes the fluctuations of the quantity $x(t)$ constitutes the second step. The CK equation yields an evolution equation for the change of the distribution function $P(x, t)$ across the scales t. The CK equation, when formulated in differential form, yields a master equation which takes the form of a Fokker-Planck equation:

$$\frac{\mathrm{d}}{\mathrm{d}t} P(x, t) = \left[-\frac{\partial}{\partial x} D^{(1)}(x, t) + \frac{\partial^2}{\partial x^2} D^{(2)}(x, t) \right] P(x, t). \tag{2}$$

The drift and diffusion coefficients, $D^{(1)}(x, t)$ and $D^{(2)}(x, t)$, are estimated directly from the data and the moments $M^{(k)}$ of the conditional probability distributions:

$$D^{(k)}(x, t) = \frac{1}{k!} \ \lim_{\Delta t \to 0} M^{(k)},$$

$$M^{(k)} = \frac{1}{\Delta t} \int \mathrm{d}x'(x' - x)^k p(x', t + \Delta t|x, t). \tag{3}$$

We note that this Fokker-Planck equation is equivalent to the following Langevin equation (8):

$$\frac{\mathrm{d}}{\mathrm{d}t} x(t) = D^{(1)}(x) + \sqrt{D^{(2)}(x)} \ f(t), \tag{4}$$

where $f(t)$ is a random force with zero mean and Gaussian statistics, δ-correlated in t, i.e., $\langle f(t) f(t') \rangle = 2\delta(t - t')$. Note that such a reconstruction of a stochastic process does *not* imply that the data do not contain any correlations, or that the above formulation ignores the correlations.

(3) Regeneration of the stochastic process constitutes the last step. Equation (4) enables us to regenerate a stochastic quantity which is similar to the original one *in the statistical sense*. The stochastic process is regenerated by iterating equation (4) which yields a series of data *without memory*. To compare the regenerated data with the original ones, we must take the spatial (or temporal) interval in the numerical discretization of equation (4) to be unity (or renormalize it to unity). However, the Markov length or time is typically greater than unity. Therefore, we should correlate the data over the Markov length or time scale. There are a number of methods to correlate the generated data in this interval [8–12]. Here, we propose a new technique which we refer to as the *kernel* method, according to which one considers a kernel function $K(u)$ that satisfies the condition that,

$$\int_{-\infty}^{\infty} K(u)\mathrm{d}u = 1, \tag{5}$$

such that the data are determined by

$$x(t) = \frac{1}{nh} \sum_{i=1}^{n} x(t_i) K \left(\frac{t - t_i}{h} \right), \tag{6}$$

where h is the window width. For example, one of the most useful kernels is the standard normal density function, $K(u) = (2\pi)^{-1/2} \exp(-\frac{1}{2}u^2)$. In essence, the kernel method represents the data as a sum of 'bumps' placed at the observation points, with its function determining the shape of the bumps, and its window width h fixing their width. It is evident that, over the scale h, the kernel method correlates the data to each other.

3 Application to fluctuations in human heartbeats

We now apply the above method to reconstruction of the fluctuations in the human heartbeats of both healthy and ill subjects by taking $h \simeq t_M$. Recent studies [13–18] reveal that under normal conditions, beat-to-beat fluctuations in the heart rate might display extended correlations of the type typically exhibited by dynamical systems far from equilibrium. It has been shown [14], for example, that the various stages of sleep may be characterized by extended correlations of heart rates separated by a large number of beats. We show that the Markov time scale t_M, and the drift and diffusion coefficients of the interbeat fluctuations of healthy subjects and patients with congestive heart failure (CHF) have completely different behaviour, when analyzed by the method we propose in this paper, hence helping one to distinguish the two groups of the subjects.

We analyze both daytime (12:00 pm to 18:00 pm) and nighttime (12:00 am to 6:00 am) heartbeat time series of healthy subjects, and the daytime records of patients with CHF. Our data base includes 10 healthy subjects (7 females and 3 males with ages between 20 and 50, and an average age of 34.3 years), and 12 subjects with CHF, with 3 females and 9 males with ages between 22 and 71, and an average age of 60.8 years). Figure 1 presents the typical data.

We first estimate the Markov time scale t_M of the data for the interbeat fluctuations. For the healthy subjects we find the Markov time scale for the daytime data
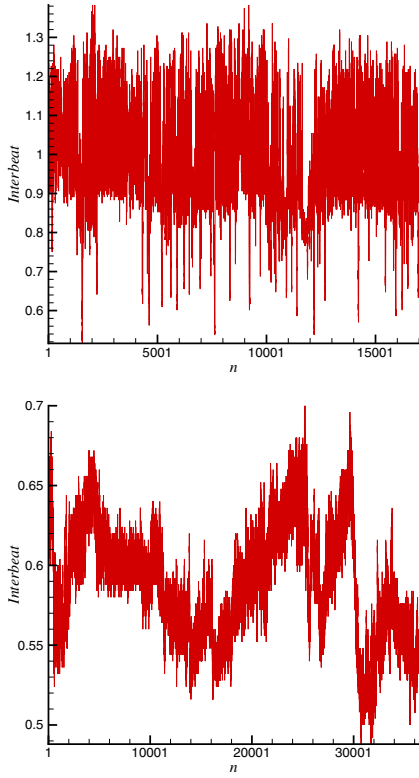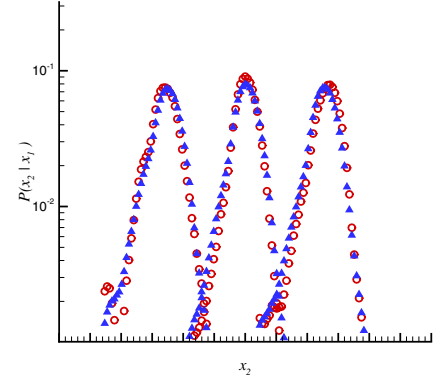
Fig. 1. Interbeats fluctuations of healthy subjects (top), and those with congestive heart failure (bottom).

Fig. 2. Test of Chapman-Kolmogorov equation for $x_1 = -5$, $x_1 = 0$ and $x_1 = 5$. The bold and open symbols represent, respectively, the directly-evaluated PDF and the integrated PDF. The PDFs are shifted in the vertical directions for better presentation. Values of x are measured in units of the standard deviation.

Fig. 3. The drift and diffusion coefficients $D^{(1)}(x)$ and $D^{(2)}(x)$, estimated by equation (3). For the healthy subjects (triangles) $D^{(1)}(x)$ and $D^{(2)}(x)$ follow linear and quadratic behavior in x, while for patients with CHF (squares) they follow third- and fourth-order behavior in x.
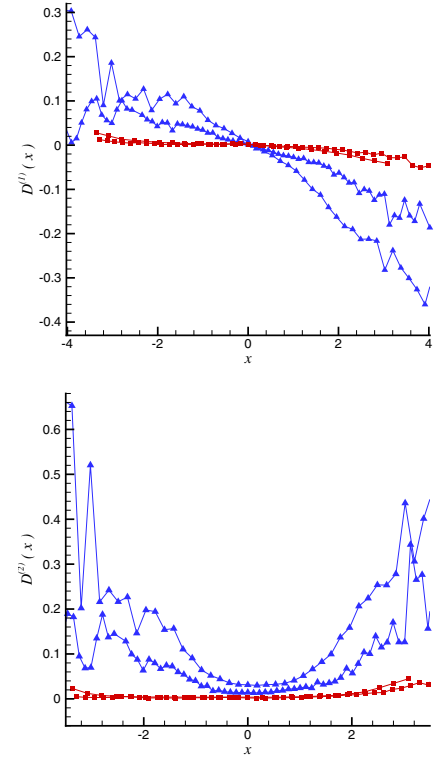
to be (all the values are measured in units of the average time scale for the beat-to-beat times of each subject), $t_M = 3, 3, 3, 1, 2, 3, 3, 2, 3$ and 2. The corresponding results for the nighttime records are, t_M are $3, 3, 1, 3, 3, 2, 3, 3, 2$ and 3, respectively, comparable to those for the daytime. On the other hand, for the daytime records of the patients with CHF, the estimated Markov time scales are, $t_M = 151, 258, 760, 542, 231, 257, 864, 8, 366, 393, 385$, and 276. Therefore, the healthy subjects have t_M values that are much smaller than those of the patients with CHF, hence providing an unambiguous quantity for distinguishing the two.

We then check the validity of the CK equation for several x_1 triplets by comparing the directly-evaluated conditional probability distributions $p(x_2, t_2|x_1, t_1)$ with the ones calculated according to right side of equation (1). Here, x is the interbeat and for all the samples we define, $x \equiv (x - \bar{x})/\sigma$, where $\bar{x}$ and σ are the mean and standard deviations of the interbeats data. In Figure 2, the two differently-computed PDFs are compared. Assuming the statistical errors to be the square root of the number of events in each bin, we find that the two PDFs are *statistically identical.*

The corresponding drift and diffusion coefficients $D^{(1)}(x)$ and $D^{(2)}(x)$ are displayed in Figure 3. We find that, in addition to the Markov time scale t_M, the two coefficients provide another important indicator for distinguishing the ill from the healthy subjects: for the healthy subjects the drift $D^{(1)}$ and the diffusion coefficient $D^{(2)}(x)$ are, respectively, a linear and a quadratic function of x,

whereas the corresponding coefficients for patients with CHF follow a third- and fourth-order equations in x. The analysis of the data yields the following approximants for the healthy subjects,

$$D^{(1)}(x) = -0.12x \ ,$$

$$D^{(2)}(x) = 0.05 - 0.042x + 0.07x^2 \ , \tag{7}$$

whereas for the patients with CHF we find that,

$$D^{(1)}(x) = -0.0026x - 0.0018x^2 - 0.0007x^3 \ ,$$

$$D^{(2)}(x) = 0.0006 - 0.0007x + 0.0005x^2$$

$$+0.0003x^3 + 0.0002x^4 \ . \tag{8}$$

Equations (7) and (8) present the drift and diffusion coefficients for a typical healthy subject and one with CHF. We note that the final result for the Langevin equation is the same as the results obtained in reference [18]. For other data measured for other patients the functional dependence of $D^{(1)}$ and $D^{(2)}(x)$ would be the same but with different numerical coefficients. The order of magnitude of the coefficients is the same for all the healthy subjects, and likewise for those with CHF (see also Ref. [19]). Moreover, if we analyze different parts of the time series separately, we find, (1) almost the same Markov time scale for different parts of the time series, but with some differences in the numerical values of the drift and diffusion coefficients, and (2) that the drift and diffusion coefficients for different parts of the time series have the same *functional forms*, but with *different coefficients* in equations such as (7) and (8). Hence, one can distinguish the data for sleeping times from those for when the subjects are awake [20].

We also find another important difference between the heartbeat dynamics of the two classes of subjects: compared with the healthy subjects, the drift and diffusion coefficients for the patients with CHF are very small (reflecting, in some sense, the large Markov time scale t_M). Hence, we suggest that one may use the Markov time scales, the dependence of the drift and diffusion coefficients on x, as well as their comparative magnitudes, for characterizing the dynamics of human heartbeats and their fluctuations, and to distinguish healthy subjects from those with CHF. To our knowledge, this proposal is novel. Given its relative simplicity, it would be most interesting to study whether this proposal can be developed into a diagnostic tool for early detection of congestive heart failure. Work in this direction is in progress.

We compare in Figure 4 the original time series $x(n)$ with those reconstructed by the Langevin equation (by, for example, using Eqs. (4) and (7)) and the kernel method. While both methods generate series that look similar to the original data, the kernel method appears to better mimic the behavior of the original data. To demonstrate the accuracy of equation (6), we compare in Figure 5 the second moment of the stochastic function, $C_2(m) = \langle [x(0) - x(m)]^2 \rangle$, for both the measured and reconstructed data using the kernel method. The agreement between the two is excellent. However, it is well-known that the agreement between the second moments of a stochastic time series and its reconstructed version is not sufficient for proving the accuracy of the reconstruction method. Hence, we have also checked the accuracy of the higher-order structure function, $S_n = \langle |x(t_1) - x(t_2)|^n \rangle$ [21]. We find that the agreement between S_n for the original and reconstructed time series for $n \leq 5$ is excellent, while the difference between higher-order moments of the two times
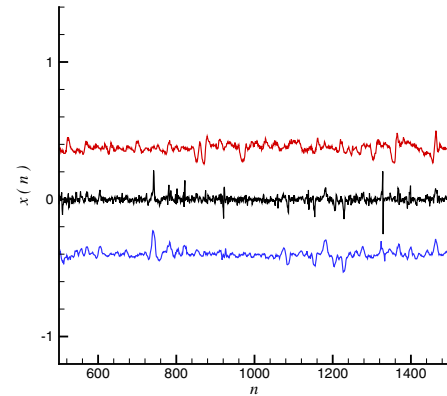
Fig. 4. The curves show, from top to bottom, the actual interbeat data (for a healthy subject), the regenerated data using the corresponding Langevin equation, and the regenerated data using the kernel method. The time series are shifted in the vertical directions for better presentation.
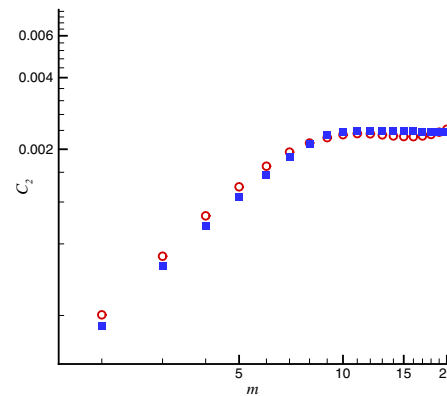
Fig. 5. Logarithmic plot of the second moment of the height-difference versus m, for the actual data (circles) and the samples regenerated by the kernel method (squares). The corresponding time series are plotted in Figure 4.

series, which are related to the tails of the PDF of the x−increments, increases.

4 Summary

We have analyzed the interbeat fluctuations in the heart rates of healthy subjects, as well as those with congestive heart failure, by an inverse method for reconstruction of the stochastic process that governs the fluctuations. The method, which is quite general and can regenerate a stochastic process with high precision, is based on utilizing measured data to estimate a drift and a diffusion coefficients to be used in a Fokker-Planck, or an equivalent Langevin, equation that describes the stochastic process. The analysis of the times series for human heartbeat dynamics using the new method, for both healthy subjects and those with CHF, not only demonstrates the accuracy of the method, but also potentially provides a novel technique for distinguishing the heartbeat dynamics of the two classes of subjects.

We should point out that Stanley and co-workers [13, 15–17, 20, 21] analyze the type of data we

considered in this paper by a method different from what we present in the present paper. Their analysis indicates that there may be long-range correlations in the data, which might be characterized by self-affine fractal distributions, such as the fractional Brownian motion or other types of stochastic processes that give rise to such correlations. They distinguish healthy subjects from those with CHF in terms of the type of correlations that might exist in the data (negative as opposed to positive correlations). The method proposed in the present paper is different from that of Stanley and co-workers in that, we analyze the data in terms of Markov processes. Although our analysis does indicate the existence of correlations in the data but, as is well-known in the theory of Markov processes, such correlations, though extended, eventually decay. We distinguish the healthy subjects from those with CHF in terms of the differences between the drift and diffusion coefficients of the Fokker-Plank equation which, in our view, provides a clearer and more physical way of understanding the differences between the two groups of the subjects. In addition, our method provides an unambiguous way of reconstructing the data, hence providing a means to predict the behavior of the data over periods of time that are on the order of the Markov time scale. Although it remains to be tested, we believe that our method is more sensitive to small differences between the data for the two groups of the subjects and, therefore, might eventually provide a diagnostic tool for early detection of CHF in humans.

References

1. S. Torquato, *Random Heterogeneous Materials* (Springer, New York, 2002); C.L.Y. Yeong, S. Torquato, Phys. Rev. E **57**, 495 (1998); C.L.Y. Yeong, S. Torquato, Phys. Rev. E **58**, 224 (1998)
2. M. Sahimi, *Heterogeneous Materials*, Vol. II (Springer, New York, 2003)
3. R. Friedrich, J. Peinke, Phys. Rev. Lett. **78**, 863 (1997)
4. R. Friedrich, J. Peinke, C. Renner, Phys. Rev. Lett. **84**, 5224 (2000)
5. R. Friedrich, Th. Galla, A. Naert, J. Peinke, Th. Schimmel, in *A Perspective Look at Nonlinear Media*, edited by J. Parisi, S.C. Muller, W. Zimmermann (Springer, Berlin, 1997), p. 313; R. Friedrich, et al., Phys. Lett. A **271**, 217 (2000)
6. M. Siefert, A. Kittel, R. Friedrich, J. Peinke, Eur. Phys. Lett. **61**, 466 (2003); S. Kriso, et al., Phys. Lett. A **299**, 287 (2002); S. Siegert, R. Friedrich, J. Peinke, Phys. Lett. A **243**, 275 (1998)
7. H. Risken, *The Fokker-Planck Equation* (Springer, Berlin, 1984)
8. G.R. Jafari, S.M. Fazlei, F. Ghasemi, S.M. Vaez Allaei, M.R. Rahimi Tabar, A. Iraji Zad, G. Kavei, Phys. Rev. Lett. **91**, 226101 (2003)
9. C. Renner, J. Peinke, R. Friedrich, J. Fluid Mech. **433**, 383 (2001)
10. M. Ragwitz, H. Kantz, Phys. Rev. Lett. **87**, 254501 (2001)
11. R. Friedrich, C. Renner, M. Siefert, J. Peinke, Phys. Rev. Lett. **89**, 149401 (2002)
12. J. Davoudi, M.R. Rahimi Tabar, Phys. Rev. Lett. **82**, 1680 (1999)
13. C.-K. Peng, J. Mietus, J.M. Hausdorff, S. Havlin, H.E. Stanley, A.L. Goldberger, Phys. Rev. Lett. **70**, 1343 (1993)
14. A. Bunde, S. Havlin, J.W. Kantelhardt, T. Penzel, J.-H. Peter, K. Voigt, Phys. Rev. Lett. **85**, 3736 (2000)
15. P. Bernaola-Galvan, P.Ch. Ivanov, L.N. Amaral, H.E. Stanley, Phys. Rev. Lett. **87**, 168105 (2001)
16. V. Schulte-Frohlinde, Y. Ashkenanzy, P.Ch. Ivanov, L. Glass, A.L. Goldberger, H.E. Stanley, Phys. Rev. Lett. **87**, 068104 (2001)
17. Y. Ashkenazy, P.Ch. Ivanov, Shlomo Havlin, C.-K. Peng, A.L. Goldberger, H.E. Stanley, Phys. Rev. Lett. **86**, 1900 (2001)
18. T. Kuusela, Phys. Rev. E **69**, 031916 (2004)
19. M.M. Wolf, G.A. Varigos, D. Hunt, J.G. Sloman, Med. J. Aust. **2**, 52 (1978)
20. P.Ch. Ivanov, A. Bunde, L.A.N. Amaral, S. Havlin, J. Fritsch-Yelle, R.M. Baevsky, H.E. Stanley, A.L. Goldberger, Europhys. Lett. **48**, 594 (1999)
21. P.Ch. Ivanov, L.A.N. Amaral, A.L. Goldberger, S. Havlin, M.G. Rosenblum, Z. Struzik, H.E. Stanley, Nature (London) **399**, 461 (1999)