Using the minimum description length principle for global reconstruction of dynamic systems from noisy time series

Ya. I. Molkov, D. N. Mukhin,* E. M. Loskutov, and A. M. Feigin
Institute of Applied Physics, Russian Academy of Sciences, 46 Uljanov Street, 603950 Nizhny Novgorod, Russia

G. A. Fidelin
Nizhny Novgorod State University, 23 Gagarin Avenue, 603600 Nizhny Novgorod, Russia
(Received 15 January 2009; revised manuscript received 2 July 2009; published 15 October 2009)

An alternative approach to determining embedding dimension when reconstructing dynamic systems from a *noisy* time series is proposed. The available techniques of determining embedding dimension (the false nearest-neighbor method, calculation of the correlation integral, and others) are known [H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer-Verlag, New York, 1997)] to be inefficient, even at a low noise level. The proposed approach is based on constructing a global model in the form of an artificial neural network. The required amount of neurons and the embedding dimension are chosen so that the description length should be minimal. The considered approach is shown to be appreciably less sensitive to the level and *origin* of noise, which makes it also a useful tool for determining embedding dimension when constructing *stochastic* models.

DOI: 10.1103/PhysRevE.80.046207 PACS number(s): 05.45.Tp, 05.40.Ca, 89.75.Hc, 95.75.Wx

I. INTRODUCTION

Methods of solution of inverse problems of dynamic system (DS) reconstruction based on the observed processes (time series, TS) generated by these systems were developed in a great number of works in the past thirty years (see, for instance [1–3], and the references therein). The interest in reconstructing deterministic dynamic systems from time series is easily explained: No complete *a priori* information about the processes running in the system is required because the first-principles models (equations of motion for a medium or individual particles, equations for the field of force, radiation transport, chemical kinetics, heat and mass transfer, and others) are not constructed in this case. The mathematical model of the studied DS is constructed on the basis of direct analysis of the observed data, generally, without assumptions about the nature of the phenomenon under consideration. This potentially allows taking into account the processes poorly studied by the time of model construction. An example of inadequacy of the first-principles models is the model of the evolution of the Earth's ozone layer popular in the mid 1980s [4] that did not describe formation of the Antarctic ozone hole because of the "neglect" of the heterochemical processes running with participation of polar stratospheric cloud particles.

The available methods of reconstructing dynamic systems from time series typically include two main steps: (1) reconstruction of the system's phase variables and (2) construction of a model reproducing behavior of the system in the corresponding region of phase space.

Reconstruction of phase variables is accomplished, for example, by the method of delay coordinates [5] in the space of dimension referred to as embedding dimension. The embedding dimension should preferably be chosen to be minimum possible. In the absence of additional information about the system, the principal technique for determining embedding dimension is the false nearest-neighbor method [6] that is easily realized. Unfortunately, this method is inefficient when the observed time series contains a pronounced noise component [1], thus making it inapplicable for reconstruction of natural systems.

The basic feature of the second step—construction of a model from time series—is the fact that it is ill-posed. Namely, there always exist an infinite number of solutions approximating the observed data with preset accuracy. It is intuitively clear that for the great majority of applications, the model will be the better the simpler it is. Widely used tools for optimal model selection are known as Bayesian information criterion (BIC) [7] and Akaike information criterion (AIC) [8]. These criteria were obtained for certain classes of statistical models of stochastic processes. However, they appear useless in the case of reconstruction of dynamical systems from *noisy* data, as will be shown below on some model examples. The authors of [9] proposed to use description length as a measure of simplicity of the model. The principle of minimum description length implies that the model corresponding to the least description length is the best. As was demonstrated in [10], this provides an effective tool for choosing *technical* parameters of the model, including the optimal number of such parameters.

In the current work, we use the principle of minimum description length (MDL) for determining embedding dimension. For this, we take the universal model in the form of an artificial neural network that includes embedding dimension as a parameter. The specific feature of using neural networks is the need to apply *physically based* prior restrictions on network parameters; hence, we generalize the definition of the description length for this case. Besides, we present the MDL invariance requirement relative to arbitrary smooth transformations of model parameters. This requirement enables, in particular, finding an explicit expression for MDL, thus simplifying the use of the MDL principle significantly.

The paper comprises two parts. In the first part, the invariant MDL form is derived and the form of the model is speci-

*mukhin@appl.sci-nnov.ru

 ©2009 The American Physical Society

fied. In the second part, it is demonstrated that, in the presence of noise when the standard methods of determining dimension are inefficient, the MDL principle allows successful solution of the problem. Besides, successful application of MDL principle to time series measured in real experiment is demonstrated.

II. DYNAMIC SYSTEM RECONSTRUCTION FROM TIME SERIES AS A PROBLEM OF OPTIMAL INFORMATION PACKING

Description length. In terms of description length minimization, data compression is an applied aspect of modeling. The result of model construction is finding the functional relationship between data points (TS elements). This allows transmitting, instead of a great number of data points, only parameters of this relationship, as well as residuals which permit the receiving party to reconstruct the data with preset accuracy. The information needed for transmitting these residuals will be referred to as data length L, and for the parameters of the relationship as model length K. The full description length is a sum of these quantities

$$F = K + L. \tag{1}$$

Data length. Following [9], the data description length will be understood as the amount of information needed for transmitting time series via some communication line. It is assumed that both the transmitter and the receiver possess some prior information about the transmitted data specified by their prior distributions. The better these distributions correspond to the real situation, the smaller the information needed for transmission is.

Let us have at our disposal a scalar time series $\{x_i\}_{i=1}^N$. We assume for simplicity that the TS is centered and normalized, i.e., $\langle x_i \rangle_i = 0$, $\langle x_i^2 \rangle_i = 1$. The relationship between the data points (global model) will be constructed in the form

$$x_i = f(x_{i-1}, \ldots, x_{i-D}, \boldsymbol{\mu}) + \xi_i, \quad \boldsymbol{\mu} \in \mathbb{R}^M. \tag{2}$$

Here, D is embedding dimension, $\boldsymbol{\mu}$ is the vector of model parameters, and M is the number of the parameters. The residuals ξ_i are supposed to be independent and normally distributed with zero mean

$$p_\xi(\xi_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\xi_i^2}{2\sigma^2}\right),$$

$$\sigma^2(\boldsymbol{\mu}) = \frac{1}{N}\sum_{i=1}^{N} [x_i - f(x_{i-1}, \ldots, x_{i-D}, \boldsymbol{\mu})]^2, \tag{3}$$

where σ is root mean square error for model data approximation.

In accordance with [9], transmission of the residual ξ_i to an accuracy of ε requires $-\ln[\varepsilon p_\xi(\xi_i)]$ nats of information. Summation over all the points yields the following estimation of data length:

$$L = -\sum_{i=1}^{N} \ln[\varepsilon p_\xi(\xi_i)] = \frac{N}{2}\left(\ln\frac{2\pi\sigma^2}{\varepsilon^2} + 1\right). \tag{4}$$

One can readily see that the data length is monotonically related to root mean square error. By increasing the dimension of the model and the number of the parameters entering it we achieve better data approximation and, hence, smaller length of their description.

Model length. Let model parameters have prior distribution $p_\mu(\boldsymbol{\mu})$. Then, the model length (or the amount of information needed for transmission of its parameters) will be

$$K = -\ln[\delta\boldsymbol{\mu} p_\mu(\boldsymbol{\mu})], \tag{5}$$

where $\delta\boldsymbol{\mu}$ is the volume in the space of parameters determining accuracy of their transmission. Assume the parameters of the model to be independent *a priori* and normally distributed with zero mean

$$p_\mu(\boldsymbol{\mu}) = \prod_{k=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{\mu_k^2}{2\sigma_k}\right), \tag{6}$$

where σ_k^2 is the dispersion of the corresponding parameter. Let us substitute Eq. (6) into Eq. (5) and write the model description length in the form

$$K = \sum_k \left[\frac{1}{2}\ln(2\pi\sigma_k^2) + \frac{\mu_k^2}{2\sigma_k^2}\right] - \ln\,\delta\boldsymbol{\mu}. \tag{7}$$

Minimum description length. Let us represent the total de-

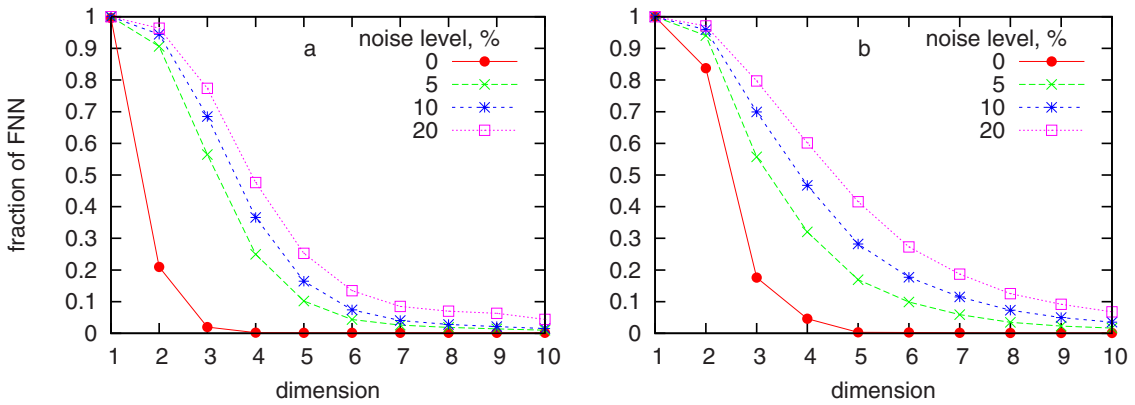

FIG. 1. (Color online) Fraction of false neighbors series for different magnitudes of noise. (a) Lorenz system; (b) Mackey-Glass system.

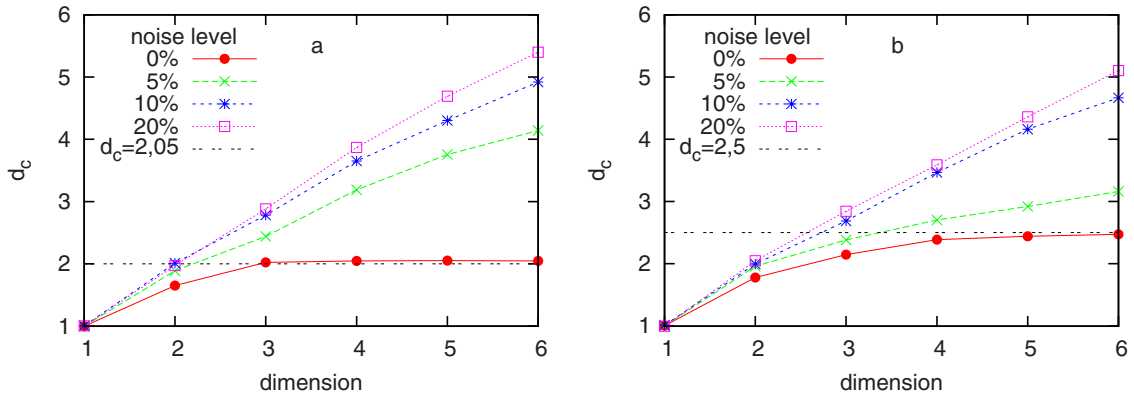

FIG. 2. (Color online) Correlation dimension versus embedding dimension. (a) Lorenz system; (b) Mackey-Glass system.

scriptive length [see Eqs. (1), (4), and (7)] in the form

$$F = \sup_{\mu \in \delta\mu} \Phi(\boldsymbol{\mu}) - \ln \delta\boldsymbol{\mu},$$

$$\Phi(\boldsymbol{\mu}) = \frac{N}{2}\left(\ln\frac{2\pi\sigma^2(\boldsymbol{\mu})}{\varepsilon^2} + 1\right) + \sum_{k=1}^{M}\left(\frac{1}{2}\ln(2\pi\sigma_k^2) + \frac{\mu_k^2}{2\sigma_k^2}\right). \tag{8}$$

The problem of seeking its minimum will be to find the optimal region $\delta\boldsymbol{\mu}$ in the space of parameters the border of which will evidently be the level surface of the function $\Phi(\boldsymbol{\mu})$. Let us perform series expansion of $\Phi(\boldsymbol{\mu})$ in the vicinity of the minimum to an accuracy of quadratic terms

$$\Phi(\boldsymbol{\mu}_0 + \boldsymbol{\delta}) = \Phi(\boldsymbol{\mu}_0) + \frac{1}{2}\boldsymbol{\delta}^T Q \boldsymbol{\delta},$$

where $Q_{ij} = \frac{\partial^2\Phi}{\partial\mu_i\partial\mu_j}\big|_{\boldsymbol{\mu}_0}$ is the second derivative matrix in the minimum. In this approximation, the level surface will be an ellipsoid oriented along the eigenvectors of matrix Q. Furthermore, we pass over to the proper basis of matrix Q, in which the description length takes on the form

$$F = \Phi(\boldsymbol{\mu}_0) + \frac{1}{2}\sum_{k=1}^{M}\delta_k^2\lambda_k - \ln\prod_{k=1}^{M}|\delta_k|. \tag{9}$$

Here, λ_k are eigenvalues of matrix Q, and the volume in the space of parameters is defined invariantly as $\delta\boldsymbol{\mu} = \Pi_{k=1}^{M}|\delta_k|$. From the condition of F minimum we obtain $\delta_k = \frac{1}{\sqrt{\lambda_k}}$, which on substitution into Eq. (9) gives MDL

$$F^* = \Phi(\boldsymbol{\mu}_0) + \frac{M}{2} + \frac{1}{2}\ln|Q|, \tag{10}$$

where $|Q|$ is understood as determinant of matrix Q.

Apparently, if complication of the model (an increase in the number of its parameters M) does not lead to a pronounced decrease in root mean square error (and, hence, to a marked decrease in data length), the MDL of the model will increase. Therefore, it is to be expected that there will be a minimum in the MDL dependence on the number of parameters.

Model. We take as a model an artificial neural network [11] in the form of a three-layer perceptron [12]

$$f(x_{k-1}, \ldots, x_{k-D}, \boldsymbol{\mu}) = \sum_{i=1}^{m}\alpha_i \tanh\left(\sum_{j=1}^{D} w_{ij}x_{j-k} + \gamma_i\right). \tag{11}$$

It was shown in [12] that this function is suitable for approximation of any regular function with preset accuracy. The

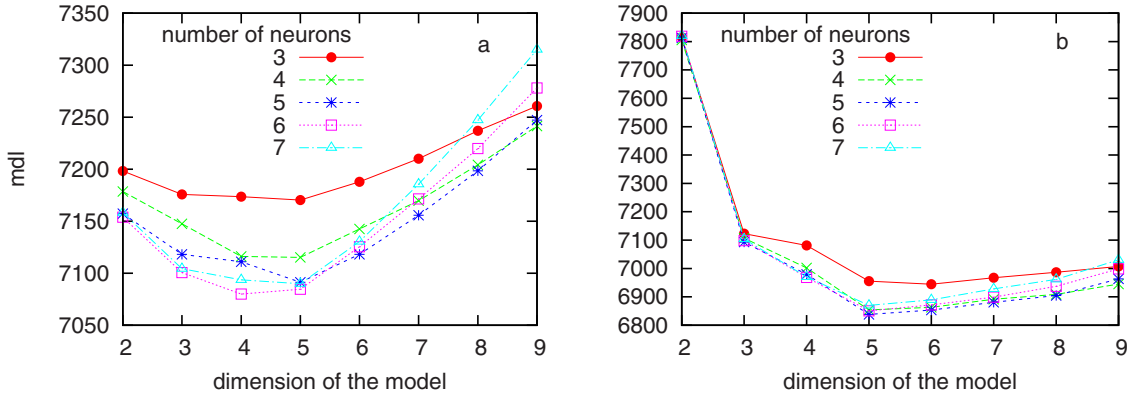

FIG. 3. (Color online) MDL versus embedding dimension for different number of neurons. Time series with 10% noise are used. (a) Lorenz system; (b) Mackey-Glass system

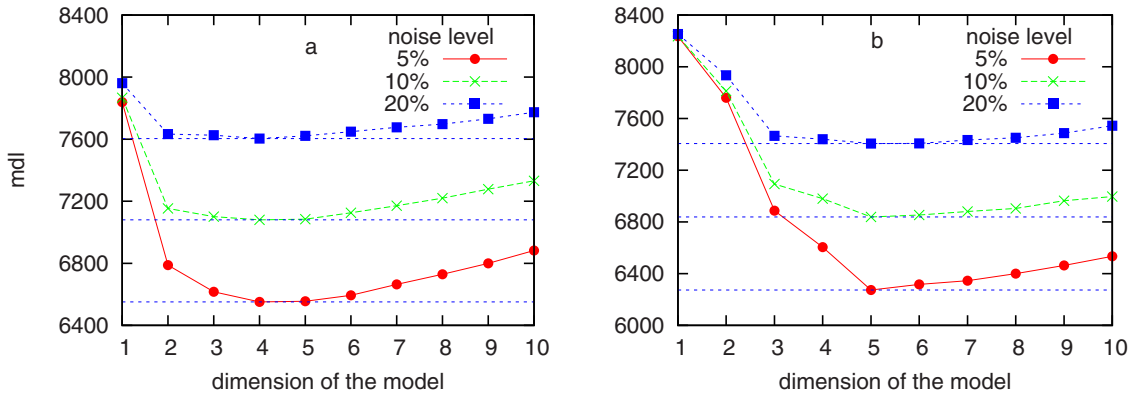

FIG. 4. (Color online) MDL versus embedding dimension for different magnitudes of noise and optimal number of neurons. Horizontal dashed lines show the minimal values of MDL. (a) Lorenz system; (b) Mackey-Glass system.

choice of such a parameterization is connected with a convenient way of defining priors for its parameters (see the next paragraph for detail). In Eq. (11) $\boldsymbol{\mu}=\{\boldsymbol{\alpha},\mathbf{w},\boldsymbol{\gamma}\}$ are network parameters, m is the number of neurons in the hidden layer, and $M=m(D+2)$ is the total number of parameters. According to [13], prior distributions of network parameters were regarded to be normal with zero mean. Output layer parameter dispersion was $\sigma_\alpha^2=\langle x_i^2\rangle/m=1/m$.

The choice of dispersion of the other parameters of these distributions is not a trivial task. In view of the chaotic nature of the time series, initially closed trajectories scatter in the phase space of the system. This means that the maximum magnitude of the derivative $|\partial f/\partial x_{k-j}|$ that determines dispersion σ_w^2 will depend on index j as $\sigma_{wj}=\exp(\lambda j)$, where λ is the maximum Lyapunov exponent. The time lag is taken to be 1.

Finally, we have

$$\sigma_\gamma^2=\langle x_i^2\rangle\sum_{j=1}^{D}\sigma_{wj}^2=e^{2\lambda}\frac{e^{2D\lambda}-1}{e^{2\lambda}-1}\approx e^{2D\lambda}.$$

III. USING MDL FOR DETERMINING MINIMUM EMBEDDING DIMENSION

In this section, we will demonstrate robustness of the MDL criterion for determining embedding dimension on examples of two broadly known systems. In the first example the Lorenz system was used [14] that is a set of three ordinary differential equations

$$\dot{x}=\sigma(y-x)$$

$$\dot{y}=x(r-z)-y$$

$$\dot{z}=xy-bz. \tag{12}$$

In our numerical experiment, we used a time series of variable x 1000-points long, sampled with the lag of 0.2 [20], generated by the system for the parameters $r=28$, $\sigma=10$, $b=8/3$ providing chaotic regime of behavior. Another series was generated by the Mackey-Glass system [15] described by the delayed differential equation

$$\dot{x}=-0.1x+0.2\frac{x(t-\tau)}{1+x(t-\tau)^{10}} \tag{13}$$

Since this system has an infinite number of degrees of freedom, it can be considered as a model of space-distributed dynamical system. A series of 1000 points, sampled with the lag 10 with parameter $\tau=23$ was used.

The generally accepted software for analysis of time series is TISEAN [16]. We present the results of assessing embedding dimension by means of two well-known methods from this software: the false nearest-neighbors method and the method of estimation of correlation dimension of an attractor. The first method gives the dependence of false neighbors fraction on dimension; this value tends to zero for true embedding dimension. Evaluation of correlation dimension by the second method allows us to plot the correlation dimension value as a function of embedding dimension; at correct values of embedding dimension one can expect a plateau. Results of application of these methods to the time series described above are shown in Figs. 1 and 2 together with results of analogous calculations for the series with 5%, 10%, and 20% measurement noise [21]. It follows from the figures that (a) the minimum embedding dimension esti-

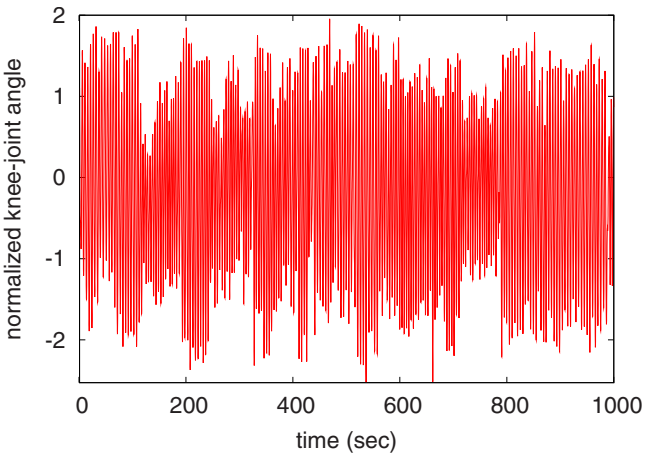

FIG. 5. (Color online) The analyzed time series of knee-joint angle. Normalized values of the angle are shown.

TABLE I. Values of optimal embedding dimension estimated by different information criteria from the analyzed time series with different magnitudes of measurement noise.

	MDL			AIC			BIC		
Model/noise level	5%	10%	20%	5%	10%	20%	5%	10%	20%
Lorenz	4–5	4	4	6	8	>10	5–6	6	6
Mackey-Glass	5	5	5–6	5–6	8	>10	5–6	7–8	7–8

mated by a noiseless series is 4 for system (12) and 5 for system (13), (b) in the presence of noise it is impossible to estimate embedding dimension by either method.

We calculated the minimum description length by the same time series. Minimization of the function (10) over $\boldsymbol{\mu}_0$ was done by the variable metric method [17]. MDL values obtained by the series with 10% noise is shown in Figs. 3(a) and 3(b) as a function of an embedding dimension for different number of neurons for both, system (12) and system (13), correspondingly. Clearly, these dependences have well-pronounced minima and there is an optimal number of neurons equal to 6 for the Lorenz system and 5 for the Mackey-Glass system. MDL versus embedding dimension is plotted in Figs. 4(a) and 4(b) for the optimal number of neurons and different magnitudes of noise.

All the dependences clearly feature the minimum embedding dimension equal to 4 for system (12) and 5 for system (13), which agrees with the results of studies by both the false nearest neighbors and correlation dimension estimator in the absence of noise.

Let us now compare the obtained values of embedding dimension with estimates given by broadly used information criteria such as AIC [8] and BIC [7]. In terms of the current paper, these criteria consist in minimization of functions $N \ln\frac{\sigma(\boldsymbol{\mu})^2}{\varepsilon^2}+2M$ and $N \ln\frac{\sigma(\boldsymbol{\mu})^2}{\varepsilon^2}+M \ln N$ over $\boldsymbol{\mu}$ and M for AIC and BIC, correspondingly. Regarding our problem parameters number M depends on both dimension D and number of neurons m, therefore we have to find the values of D and m minimizing these functions. We calculated these values for our time series; the corresponding results are shown in Table I. It is clear from this table that the estimates of dimension given by AIC and BIC grow with increasing noise level for both the considered systems, while the same estimates given by MDL principle remain approximately fixed.

Finally, we consider the application of MDL principle to time series of locomotory motions of humans. We used results of the experiment described in [18], in which automatic stepping in healthy humans under appropriate afferent stimulation was investigated. In paper [19], the authors hypothesized existence of central rhythm generator controlling such a motion. They created the phenomenological dynamical model including four first-order differential equations, which demonstrates behavior with similar dynamic properties as the observed processes. In particular, it reproduces spontaneous switchings between dynamical regimes with backward and forward steps occurring in the experiment. In other words, it was shown that the model with dimension d_m equal to approximately (at least) 4 can be sufficient for reconstruction of basic dynamical properties of the system underlying the observed dynamics. This means that the embedding dimension of the attractor reconstructed from single time series by the delay method is expected to be less than $2d_m+1=9$. Now, we will try to determine the optimal dimension for construction of a global model using the MDL principle. The time series we use is the oscillogram of knee-joint angle (see Fig. 5). The dependence of MDL on embedding dimension is shown in Fig. 6(a) for different numbers of neurons. It is clear from this figure that the optimal number of neurons is 4–5 and the optimal embedding dimension is equal to 5–7, which is in good agreement with the conclusion given in [19]. The estimate of rms of noise component of time series is defined by the value of $\sigma(\boldsymbol{\mu})$ at $\boldsymbol{\mu}$ corresponding to the minimum description length. After division of the time series into rms, it is equal to 0.15 for optimal number of neurons and optimal dimension. The result of false nearest-neighbors method ap-

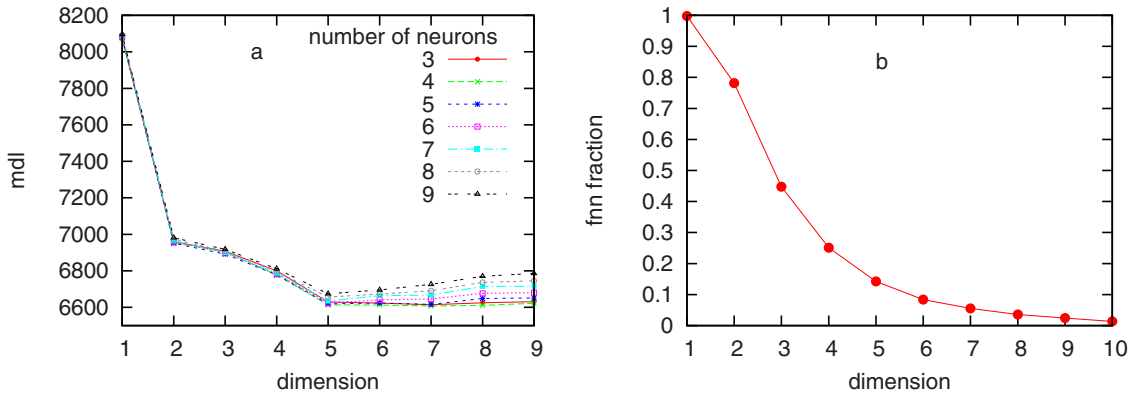

FIG. 6. (Color online) Analysis of human locomotion data (a) MDL versus embedding dimension for different number of neurons. (b) Fraction of false nearest-neighbors versus dimension.

plication is shown in Fig. 6(b) for comparison; it is clear that this method is useless in the considered situation. The latter is most likely connected with a sufficiently high noise level in the analyzed data.

IV. CONCLUSION

Although methods of dynamic system reconstruction from time series have been considered in a great amount of papers, there are only scant examples of their useful application to "natural" data obtained in experiments. One of the possible reasons is the presence in experimental time series of a noise component, generally of unknown origin. In the presence of noise, the available techniques cannot specify the dimension for reconstruction, even when the observed process may be regarded to be the product of the evolution of a deterministic dynamic system. The MDL method is more robust to the magnitude of noise than the false nearest neighbors and other existing techniques. Hence, it may be used in the case under consideration.

It is also worthy of notice that the form of model (2) corresponds to the case when noise in the system is dynamic. Thus, the MDL principle may be used for reconstruction of *stochastic* systems from the time series generated by them. Reconstruction of stochastic systems will be considered elsewhere.

ACKNOWLEDGMENTS

The work was supported by the Russian Foundation for Basic Research (Project No. 06-02-16568) and by the Program of Basic Research of the RAS Presidium Fundamental Problems of Nonlinear Dynamics (Project No. 3.3)

[1] H. D. I. Abarbanel, *Analysis of Observed Chaotic Data* (Springer-Verlag, New York, 1997).
[2] B. Bezruchko and D. Smirnov, *Mathematical Simulation and Chaotic Time Series* (College, Saratov, 2005) (in Russian).
[3] V. Anishchenko, T. Vadivasova, and V. Astakhov, *Nonlinear Dynamics of Chaotic and Stochastic Systems* (Saratov University, Saratov, 1999) (in Russian).
[4] *Final Report National Academy of Sciences–National Research Council, Washington, D.C., Commission on Physical Sciences, Mathematics and Resources* (National Academy of Sciences–National Research Council, Washington, D.C., 1984).
[5] F. Takens, Lect. Notes Math. **898**, 366 (1981).
[6] M. B. Kennel, R. Brown, and H. D. I. Abarbanel, Phys. Rev. A **45**, 3403 (1992).
[7] G. E. Schwarz, Ann. Stat. **6**, 461 (1978).
[8] H. Akaike, IEEE Trans. Autom. Control **19**, 716 (1974).
[9] K. Judd and A. Mees, Physica D **82**, 426 (1995).
[10] Z. Yi and M. Small, IEEE Trans. Circuits Syst., I: Regul. Pap. **53**, 722 (2006).
[11] K. Hornik, M. Stinchcombe, and H. White, Neural Networks **2**, 359 (1989).
[12] *The Handbook of Brain Theory and Neural Networks*, edited by M. A. Arbib (MIT, Cambridge, MA, 1995).
[13] R. M. Neal, Department of Computer Science, University of Toronto Technical, Report No., CRG-TR-93-1, 1993 (unpublished).
[14] E. N. Lorenz, J. Atmos. Sci. **20**, 130 (1963).
[15] M. C. Mackey and L. Glass, Science **197**, 287 (1977).
[16] H. Kantz and T. Schreiber, *Nonlinear Time Series Analysis* (Cambridge University Press, Cambridge, England, 1997).
[17] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C* (Cambridge University Press, Cambridge, 1992).
[18] V. Gurfinkel, Y. Levik, O. Kazennikov, and V. Selionov, Eur. J. Neurosci. **10**, 1608 (1998).
[19] A. K. Kozlov, M. M. Sushchik, Ya. I. Molkov, and A. S. Kuznetsov, Int. J. Bifurcation Chaos Appl. Sci. Eng. **9**, 2271 (1999).
[20] The problem of choice of an optimal time lag for phase variables reconstruction was discussed in many works (see, for instance, Refs. [1,16]). We use time lag corresponding to the minimum of mutual information function (Ref. [16].
[21] Gaussian white noise is used; hereinafter the noise level is defined as ratio of noise and data rms.