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

