Journal of Atmospheric and Solar-Terrestrial Physics 140 (2016) 114–119

Contents lists available at ScienceDirect

Journal of Atmospheric and Solar-Terrestrial Physics

journal homepage: www.elsevier.com/locate/jastp

Research paper

The minimal database size and resolution of the locally linear algorithm of direct dependence recovery in helio-biology studies

V.A. Ozheredov, T.K. Breus *

Space Research Institute Russian Academy of Sciences, Moscow, Russia

ARTICLE INFO

Article history:
Received 19 August 2015
Received in revised form
30 December 2015
Accepted 4 February 2016
Available online 10 February 2016

Keywords:
Multidimensional functional relationship
Noise
Noise reduction
Local approximation
Resolution
Helio-biology

ABSTRACT

Several problems can emerge in front of investigators, who take a detailed restoration of dependency. The key of them – is a mathematically rigorous formulation of the desired degree of details. Second in importance is the reliability problem of reproduction of these details. And the third problem is the evaluation of data collection efforts that will ensure the desired depending on the required details and results reliability. In this work the strict concept of spatial resolution of the locally linear algorithm of direct dependence recovery (DDR) is formulated mathematically. Such approach implies approximation of the system reaction (dependent variable) in the case of the assigned value of factors which only utilizes the data (precedents) from a spherical cluster surrounding those assigned value of factors. The concept of reliability of details is formalized through the noise attenuation coefficient. We derive a relationship between the size of the minimum required database, spatial resolution of the recovery algorithm, the number of influencing factors and the noise attenuation coefficient. Analytical findings are verified by numerical experiments. Maximum number of factors, functional dependence on which can be recovered via the database figuring in various helio-biological works published by many authors for several 10 of years, is estimated. It is shown that the minimum required size of the database depends on the number of influencing factors (dimension of space of the independent variable) as a power law. The analysis conducted in this study reveals that the majority of the dimensional potentials of helio-biological databases are significantly higher that dimensions, which are appear in the approaches of authors of these works.

1. Introduction

For decades the effects of weather and helio-geomagnetic parameters on a human being are a subject of numerous studies in many countries. The researches usually focused on functional dependence of physiological characteristics on solar activity factors (see, for example, review paper by Palmer et al. (2006)). The vast majority of the researchers employ correlation analysis because (a) there is great choice of standard software, which is not demanding specific mathematical skills, and (b) hardware requirements are rather modest. However, the standard correlation analysis implies approximately linear dependence of the analyzed physiological characteristics on solar activity factors. In the papers by numerous authors (for example, Villoresi et al. (1994), Persinger and Psych (1995), Mitsutake et al. (2005), Gavish et al. (2008), Breus et al. (2008, 2010), Zenchenko (2010), Stoupel et al. (2007, 2011), Yu et al. (2012) and Zeng et al. (2013); and other papers listed in the reviews by Vladimirski and Temyuriants (2000), Cherry (2002) and Palmer et al. (2006)) the linear methods are used for more than 20 years. The linear dependence is just a particular case of more general non-linear one. In the absence of a priori knowledge regarding real dependence analyzed it would be reasonable to utilize more general class of non-linear models. Recalling that in this paper we focus on non-linear models.

A data base, which is obtained as a result of helio-biophysical process studies, comprises a number of dependent variable (adaptor), which dependence is to be studied, and corresponding independent parameter (predictor). The adaptor and predictors comprise a precedent. Number of independent parameters, which are taken into account, is the dimension of predictor space. The size of database ranges from several hundred to several millions of precedents while dimensions of predictor space are from 1 to 3 (see, for example, Stoupel et al. (2007, 2011), Zenchenko et al. (2013) and Zeng et al. (2013)).

Naturally, there is some noise in the adaptor response to the predictors, which is caused by the processes ignored in the study.

* Corresponding author.
E-mail addresses: ojymail@mail.ru (V.A. Ozheredov), breus36@mail.ru (T.K. Breus).

http://dx.doi.org/10.1016/j.jastp.2016.02.005
1364-6826/

The ratio of the mean square root of noise amplitude to the approximation error is the noise attenuation coefficient. The ratio of the size of the predictor space region, where the precedents are located, to minimal size of the detail, which has to be resolved, is the resolution of the algorithm for deriving functional dependence of an adaptor on predictor.

In the present paper the problem of accurate evaluation of minimal size of database required for assigned resolution, noise attenuation coefficient and dimension of predictor space (number of independent parameters, which are taken into account) is solved. This makes also possible accurate evaluation of maximal dimensions of the predictor space, which is admitted by the already available database (the set of adaptors), when resolution and noise attenuation coefficient are specified. Such approach of the database size evaluation depending on following parameters of the model: (a) the dimension of the predictor space, (b) the resolution and (c) the noise attenuation coefficient is new and very useful for helio-biology.

2. Introduction to the direct recovery of functional dependence

Problems of forecasting a system reaction to some set of external factors assumes existence of a decision enabling system, which generates expected value of the adaptor η in response to an assigned set of predictor values (then termed as forward-predictor $\hat{\boldsymbol{x}}$).

In the case under consideration, the adaptor η is assumed to be a sum of (a) a regular part $\hat{\boldsymbol{y}}=\boldsymbol{y}(\boldsymbol{x})$, which is functionally dependent on value of predictor $\boldsymbol{x}$, which is assumed to absolutely accurately be determined by predictor $\boldsymbol{x}$; and (b) a noise ν , which is centered at expectation, stationary and delta correlated. Then, the value of $\hat{\boldsymbol{y}}$ is assumed to be a reaction to the value of predictor $\boldsymbol{x}$.

The algorithm for deriving $\hat{\boldsymbol{y}}$ corresponding to any admitted value of predictor $\boldsymbol{x}$ is similar to the problem of noise filtering. In the focus of present paper there are methods of direct dependence recovery (DDR) (Ozheredov et al., 2010). The DDR methods does not utilize any a priory assumptions ("physical considerations") and, thus, only use a posterior information regarding relationship of reaction and cause. All the required information is coded in the precedent database (PR), that is in array of real pairs of adaptor–predictor, which has been collected during some period of time: $\boldsymbol{PR}=[\boldsymbol{X}\boldsymbol{\eta}]$. Here [::] indicates horizontal concatenation of columns.

3. The nearest neighbors' method and resolution concept

As already mentioned, the DDR methods does not utilize any "physical considerations" and, thus, only the precedent database (PR) is significant. However, without any "physical" assumptions regarding reason–consequence relationships the method needs some extra limitations. The DDR methods assume that the first derivative of the function $\boldsymbol{y}(\boldsymbol{x})$ exists. In particular, it is assumed that for any forward–predictor there is some spherical neighborhood around the forward–predictor where $\boldsymbol{y}(\boldsymbol{x})$ is approximately linear. Next, within the spherical neighborhood a number of precedents comprising the database PR have to be high enough for statistical attenuation of the effect of noise ν. Let $\boldsymbol{n}$ be the minimum number of the precedents, which provides sufficient attenuation of noise effect in the case of locally linear "predictor–adaptor" relationship.

The DDR method approximates the adaptor value, which corresponds to a forward–predictor $\hat{\boldsymbol{x}}$, using $\boldsymbol{n}$ precedents from the surrounding spherical neighborhood in the predictor space – the nearest neighbors. The DDR method smooths (blurs) the variations of the original function $\boldsymbol{y}(\boldsymbol{x})$ over the sphere of $\boldsymbol{r_n}$, which is the most probable radius of a sphere surrounding the forward–predictor $\hat{\boldsymbol{x}}$ and containing $\boldsymbol{n}$ precedents from the database. Thus, $\boldsymbol{r_n}$ is the minimal scale-length of variations of $\boldsymbol{y}(\boldsymbol{x})$, which can be resolved by the DDR method.

Let the distribution of predictors in the predictor space be characterized by a stationary distribution density $\boldsymbol{p}_{\xi}(\boldsymbol{x})$ and $\boldsymbol{r}_0$ be the radius of a sphere surrounding predictor $\hat{\boldsymbol{x}}$, which contains a generated predictor with probability at least 50%. Obviously, $\boldsymbol{r}_0$ is a characteristic of possible scale-length of the predictor cloud in the predictor space.

In general, the resolution of the local approximation method can be characterized by the ratio

$$\boldsymbol{Res} = \frac{\boldsymbol{r}_0}{\boldsymbol{r_n}}. \tag{1}$$

Assuming that $\boldsymbol{p}_{\xi}(\boldsymbol{x})$ is approximately multi-dimensional normal distribution, we evaluate the nominator in Eq. (1). Then, by rescaling predictors, we come to $\boldsymbol{p}_{\xi}(\boldsymbol{x})$, which is the standard centered on expectation Gaussian distribution normalized by unit. The square of the distance between a precedent and the distribution center follows the hi-square distribution with m degrees of freedom. Utilizing Cornish–Fisher series for the inverse integral distribution function, one gets that with the probability higher than 50%, a variable following hi-square distribution with m degrees of freedom is less than $\sqrt{\boldsymbol{m}}$. Thus,

$$\boldsymbol{r}_0 = \sqrt{\boldsymbol{m}}. \tag{2}$$

For evaluating the denominator in the right-hand site (RHS) of Eq. (1) we introduce an accidental variable $\rho(\boldsymbol{n})$, which is the distance from a forward predictor to the n-th nearest neighbor in the precedent database $\boldsymbol{PR}$. Let $\boldsymbol{r}$ be the distance from the forward predictor and $\boldsymbol{dr}$ be infinitely small increment of it. The event $\{\boldsymbol{r} < \rho(\boldsymbol{n}) < \boldsymbol{r}+\boldsymbol{dr}\}$ is an equivalent of sum of all possible permutations of $\boldsymbol{N}$ precedents from $\boldsymbol{PR}$ when (a) $\boldsymbol{n}-1$ of precedents fall within $\boldsymbol{\Omega}(\boldsymbol{r})$, which is the spherical neighborhood of radius $\boldsymbol{r}$ surrounding the forward–predictor $\hat{\boldsymbol{x}}$, (b) $\boldsymbol{N}-\boldsymbol{n}$ are outside of $\boldsymbol{\Omega}(\boldsymbol{r})$ and (c) one precedent falls within a spherical layer of thickness $\boldsymbol{dr}$ adjoining $\boldsymbol{\Omega}(\boldsymbol{r})$ (see Fig. 1). Recalling uncorrelated generation of the precedents in PR, the probability of the event $\{\boldsymbol{r} < \rho(\boldsymbol{n}) < \boldsymbol{r}+\boldsymbol{dr}\}$ satisfies the relation:

$$P\{r < \rho(n) < r + dr\} = C\left(\int_{\Omega(r)} p_{\xi}(x)dx\right)^{n-1} \frac{d}{dr}\int_{\Omega(r)} p_{\xi}(x)dxdr\left(1 - \int_{\Omega(r)} p_{\xi}(x)dx\right)^{N-n}.$$

Here $\boldsymbol{C}$ is the number of the aforementioned permutations. Then, one gets the relation for the distribution density $(\boldsymbol{n})$:

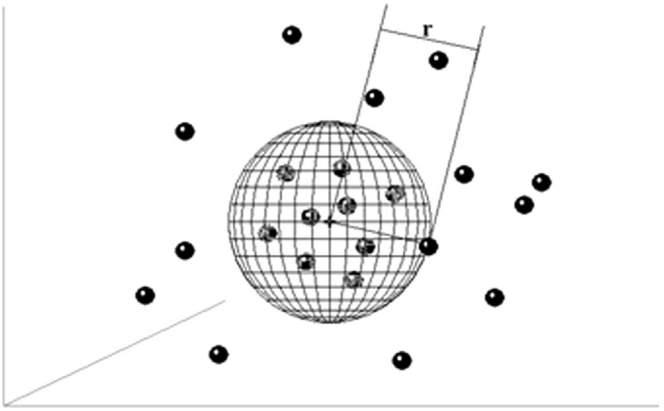

Fig. 1. An example of spatial distribution of the nearest neighbors of the forward predictor.