Physica A 419 (2015) 729–735

Contents lists available at ScienceDirect

Physica A

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

A cumulative entropy method for distribution recognition of model error

Yingjie Liang, Wen Chen *

Institute of Soft Matter Mechanics, College of Mechanics and Materials, Hohai University, No. 1 XiKang Road, Nanjing, Jiangsu 210098, China

HIGHLIGHTS

- A cumulative entropy method is proposed to recognize the distribution of model error.
- Lévy stable distribution is used to detect the statistical properties of model error.
- Compared with Shannon entropy, the cumulative entropy method is easy to be validated.
- Lévy stable distribution can well depict the distribution of model error.

ARTICLE INFO

Article history:
Received 14 August 2014
Received in revised form 25 September 2014
Available online 27 October 2014

Keywords:
Cumulative entropy
Model selection criterion
Lévy stable distributions
Shannon entropy
Model error
CFT

ABSTRACT

This paper develops a cumulative entropy method (CEM) to recognize the most suitable distribution for model error. In terms of the CEM, the Lévy stable distribution is employed to capture the statistical properties of model error. The strategies are tested on 250 experiments of axially loaded CFT steel stub columns in conjunction with the four national building codes of Japan (AIJ, 1997), China (DL/T, 1999), the Eurocode 4 (EU4, 2004), and United States (AISC, 2005). The cumulative entropy method is validated as more computationally efficient than the Shannon entropy method. Compared with the Kolmogorov–Smirnov test and root mean square deviation, the CEM provides alternative and powerful model selection criterion to recognize the most suitable distribution for the model error.

1. Introduction

The model error statistical analysis is to select the most suitable model by revealing the statistical properties of model error, which is popular in many complex random systems, such as wind power forecasts [1], circular concrete-filled steel tubular capacity design [2], and forest carbon dynamics prediction [3]. In the above-mentioned systems, the statistical properties of model error decide the design value of target variable, e.g. capacity, which can indirectly improve the reliability, safety and health of the systems, and thus the recognition of its statistical properties has become an important task. In practice, many statistical models, such as Gaussian and Lognormal distributions, can visually well-fit the distribution of model error and these candidate statistical distributions satisfy the goodness-of-fit test [2]. However, it is not easy to recognize the most suitable distribution from the candidate distributions based on the traditional model selection criteria, such as Kolmogorov–Smirnov (K–S) test and root mean square deviation (RMSD) [4]. Therefore, new model selection criterion is urgently needed to aid researchers in recognizing the most suitable distribution for model error.

* Corresponding author. Tel.: +86 25 8378 6873; fax: +86 25 8373 6860.
E-mail address: chenwen@hhu.edu.cn (W. Chen).

http://dx.doi.org/10.1016/j.physa.2014.10.077
0378-4371/

As the model error is considered to be a random system, entropy of the system can be tentatively used to recognize the most suitable distribution, which is a measure of uncertainty in a random variable [5]. Currently, the Shannon entropy, a popular concept in information theory, is most commonly used to quantify the average unpredictability in a random system [6]. However, this traditional entropy method intensively depends on the interval length of the random variable histograms with very high computational costs and low accuracy [7]. To overcome the above-mentioned troublesome issues, this study proposes a cumulative entropy method (CEM). The CEM only depends on the cumulative distribution function (CDF) of a random variable, which is more easily obtained than the probability density function (PDF), and can reflect the effects of all experimental data to quantify the uncertainty.

It is observed that the skewness and kurtosis values of model error often deviate from zero [3], while Gaussian or Lognormal distribution does not include the essential information of the tail distribution. The Lévy stable distribution [8], which can successfully fit the large data sets exhibiting heavy tails and skewness, is proposed to depict the anomalous statistical properties of such a model error. In general, the Lévy stable distribution requires four parameters: stability index α, skewness parameter β, scale parameter γ, and location parameter δ. Gaussian ($\alpha = 2$) and Cauchy distributions ($\alpha = 1$) are its two most familiar particular cases [9]. Recently, Lévy stable distributions have been applied as a statistical model to diverse systems, such as magnetic resonance imaging [10], pricing of stock [11], wind power system [12], and the noise in a communication system [13].

The rest of the paper is organized as follows. In Section 2, a cumulative entropy method is proposed to recognize the most suitable distribution from the candidate distributions, and then the cumulative distribution function of Lévy stable distributions is introduced. Section 3 examines the strategies to analyze the experimental data with four design codes of circular concrete-filled steel tubular stub columns capacity. Finally, we draw some conclusions in Section 4.

2. Methodologies

2.1. Cumulative entropy method

Cumulative entropy is derived from the mathematical form of the traditional Shannon entropy [6]. Since most experimental data of a random system are discrete, we give a discrete expression of the cumulative entropy.

$$E = -\sum_{i=1}^{n} \frac{p(x_i)}{P(x)} \log_b \frac{p(x_i)}{P(x)}, \tag{1}$$

where $p(x_i)$ is the cumulative probability of x_i, n number of the system states, i.e. the length of x_i, b the base, and $P(x) = \sum_{i=1}^{n} p(x_i)$. It can be verified that the cumulative entropy E is non-negative and expansible.

The traditional Shannon entropy of a discrete random variable x_i is defined as

$$H = -\sum_{i=1}^{n} f(x_i) \log_b f(x_i), \tag{2}$$

where $f(x_i)$ is the frequency or probability of x_i, which is obtained by the histograms of experimental data, n number of the system states, i.e. the length of sample size, b the base. However, in most cases, the probability density function (pdf) of x_i is continuous, such as Gaussian and Lognormal distributions, thus to calculate the entropy of random system, the following formula is needed to be numerically estimated.

$$H = -\int f(x_i) \log_b f(x_i), \tag{3}$$

where $f(x_i)$ is the pdf of x_i. Compared with the cumulative entropy, the Shannon entropy is more complicated.

According to formulas (1) and (2), the most distinct difference between the cumulative and the Shannon entropies is that the former is correlated with the CDF of a random system, while the latter depends on interval length of the experimental data histograms, i.e., the PDF. Compared with the Shannon entropy, the CEM is easier to be established in terms of computing costs and accuracy. In conjunction with the formula (1), the distance between the predicted cumulative entropy and the standard cumulative entropy of a random model error system is defined as ε.

$$\varepsilon = |E_{pe} - E_{se}|, \tag{4}$$

where E_{pe} is the predicted cumulative entropy, E_{se} the standard cumulative entropy. The standard cumulative entropy is calculated with the empirical cumulative probability of the model error system. The respective distance ε of the candidate distributions is estimated by the expression (4), which forms a distance set. The minimum value in the distance set indicates that its corresponding distribution model is the most suitable one among the candidate distributions.

2.2. Cumulative distribution function of Lévy stable distributions

The CDF of Lévy stable distributions has no explicit formulas except for a few particular cases such as the Gaussian and Cauchy distributions. A famous method to evaluate its CDF was proposed by Zolotarev [14], in which the following