Chemical Physics Letters 639 (2015) 36–40

Contents lists available at ScienceDirect

Chemical Physics Letters

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

Benford analysis: A useful paradigm for spectroscopic analysis

Gaurav Bhole, Abhishek Shukla, T.S. Mahesh *

Department of Physics and NMR Research Center, Indian Institute of Science Education and Research, Pune 411008, India

ARTICLE INFO

Article history:
Received 15 April 2015
In final form 27 August 2015
Available online 3 September 2015

ABSTRACT

Benford's law is a statistical inference to predict the frequency of significant digits in naturally occurring numerical databases. In such databases this law predicts a higher occurrence of the digit 1 in the most significant place and decreasing occurrences to other larger digits. Although counter-intuitive at first sight, Benford's law has seen applications in a wide variety of fields like physics, earth-science, biology, finance, etc. In this work, we have explored the use of Benford's law for various spectroscopic applications. Although, we use NMR signals as our databases, the methods described here may also be extended to other spectroscopic techniques. In particular, with the help of Benford analysis, we demonstrate emphasizing weak NMR signals and spectral corrections. We also explore a potential application of Benford analysis in the image-processing of MRI data.

1. Introduction

Benford's law is an empirical law, which applies to many of the naturally occurring numerical datasets [1]. It predicts the probability of digits 1–9 being the most significant digit (left-most nonzero digit of a nonzero number). Naively one expects that the frequency of digits 1–9 in the most significant place is nearly uniform, unless the entries in the dataset are biased to one or more digits due to a constrained range of the entries. On the contrary, Benford's law predicts a higher occurrence of the digit '1' in the most significant place, compared to other higher digits, whose occurrences decrease progressively. This non-uniform occurrence of digits in the most significant place was first observed by Simon Newcomb [2] in 1881. However, Newcomb's article failed to gain recognition due to a lack of mathematical structure. This empirical law was rediscovered by Frank Benford [1] in 1938, who presented it with a mathematical formulation, which states that in a given dataset, the probability $P_B(d)$ of the most significant digit 'd' is given by:

$$P_B(d) = \log_{10}\left(1 + \frac{1}{d}\right). \quad (1)$$

This probability distribution, known as Benford's law, predicts an occurrence as high as 30.1% for the digit '1' in the most significant place, whereas a mere 4.6% for the digit 9.

Benford distribution is often attributed to our numbering-systems (decimal, octal, or hexadecimal) which are linear unlike many natural processes which are geometric progressions [1]. For example, several phenomena such as the Maxwell–Boltzmann distribution of thermal energy, pH values of solutions, charging/discharging of capacitor, Newton's law of cooling, decay of a radioactive nuclei – all occur in exponents.

In order to provide a simple illustration of Benford's law, we consider an exponentially decaying signal. While such signals are versatile, let us consider a time-domain free-induction decay (FID) of a single NMR transition. Figure 1(a) displays such a digitized signal, and also marks the time-intervals at which digits 1–9 occur in the most significant place. The percentage of the total occurrence of each digit is also shown in the figure. Evidently, this distribution is in excellent agreement with the expression (1). Let us now consider the corresponding frequency-domain signal which is routinely used in spectroscopic analysis. Figure 1(b) displays the real part of the Fourier transform of the time-domain signal shown in Figure 1(a). Here the distributions of most significant digits indicate an overall agreement with Benford's law.

Benford distribution is not limited to only exponential or Lorentzian functions. Any unconstrained dataset whose entries range over several orders of magnitude tend to follow Benford's law. Such datasets emerge from a plethora of sources ranging from astrophysical [3], geographical [4], seismographic [5], biological [6–8], to financial topics [9–11]. Benford's law has also been applied to study phase transitions in quantum many-body systems [12,13]. Recently in another study, we had also discussed the possibility of distinguishing a genuine NMR spectrum from a simulated one using Benford analysis [14].

In this article, we explore the use of Benford's law for various spectroscopic applications, specifically in NMR. In particular, we

* Corresponding author.
E-mail address: mahesh.ts@iiserpune.ac.in (T.S. Mahesh).

http://dx.doi.org/10.1016/j.cplett.2015.08.061
0009-2614/

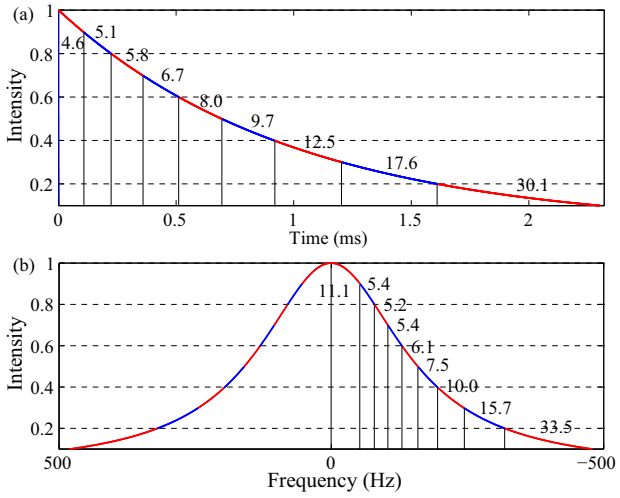

Figure 1. (a) An exponentially decaying signal and the distribution of various digits in the most significant place. (b) Similar distribution for the real part of the Fourier transform of (a). The numbers above each curve indicate the percentage of the time duration (or the frequency range) spent by each part of the curve (ignoring parts with intensities less than 0.1).

present some useful techniques pertaining to spectroscopic peak detection, phase correction, and baseline correction. Firstly, we describe the methodology of Benford analysis. Then in the next section, we demonstrate three applications in NMR spectral analysis (i) emphasizing weak solute peaks which are overridden by strong solvent peaks in 1D and 2D NMR, (ii) simultaneous correction of zeroth order and first order phase errors in NMR spectra, and (iii) a technique for baseline correction. Moreover, we show that the Benford analysis of an MRI image might be helpful in highlighting its key areas. Finally we conclude in the last section.

2. Methodology of Benford analysis

2.1. Benford goodness parameter

We shall describe here the general procedure adopted for performing Benford analysis. Firstly, the signs of the data entries $\{x_i\}$ are removed by taking their absolutes. To remove a bias towards any particular digit(s) arising from a limited data-range $[x_{\min}, x_{\max}]$, the dataset is rescaled between 0 and 1, by using the transformation [13]

$$x_i \to \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}. \tag{2}$$

The distribution $P(d)$ of digits $d = \{1, \ldots, 9\}$ in the most significant place is now extracted. In order to quantify the extent of agreement between the observed distribution $P(d)$ with the expected distribution $P_B(d)$, we define 'Benford Goodness Parameter' (BGP) as [4]

$$BGP = \left(1 - \sqrt{\sum_{d=1}^{9} \frac{(P(d) - P_B(d))^2}{P_B(d)}}\right) \times 100. \tag{3}$$

An ideal Benford distribution corresponds to a value of BGP = 100, whereas the datasets encountered in real life may take lower or even negative values.

2.2. Dependence on phase, baseline, and noise

Now we describe the dependency of BGP on certain spectral parameters which form the basis for some interesting applications

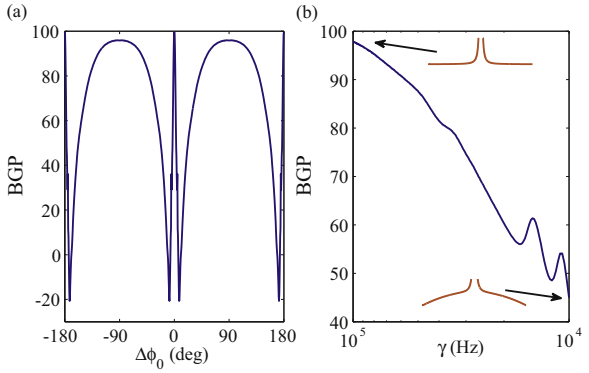

Figure 2. BGP versus (a) phase-error and (b) the baseline-error controlled by the γ parameter.

in spectroscopy: (i) The BGP values are sensitive to the spectral line-shapes and not to their relative intensities. (ii) As illustrated in Figure 2(a), the BGP value depends on the phase of the spectral line. For a given Lorentzian spectral line, maximum BGP is displayed when it is perfectly absorptive ($\phi_0 = 0$). (iii) The BGP value also depends on the baseline. This fact is illustrated in Figure 2(b). Here a Lorentzian peak is placed on a Gaussian base line, $\exp(-\nu^2/\gamma^2)$ wherein ν is the frequency and γ is the width parameter. Clearly, the BGP increases as the baseline becomes flatter.

To study the noise-dependence of BGP values, we recorded a series of ^{1}H spectra of water with varying signal to noise ratios. The results are plotted in Figure 3. BGP displays a weak overall enhancement with the signal to noise ratio. This is expected since the noise tends to alter the distribution of the Lorentzian function. Further analysis of the noise dependence has been described in [14].

2.3. Scanning Benford analysis

So far we have considered only one spectral line, while many spectral lines are encountered in a typical NMR signal. In such situations, it is helpful to consider a 'bin' formed by a smaller section of the data set and compute its BGP. A set of BGP values are then obtained after systematically moving the bin over the entire dataset. We can then plot the BGP values versus the bin centers. We refer to this procedure as Scanning Benford analysis (SBA). The mean of the BGP values thus obtained captures the lineshapes at all parts of the spectrum, and therefore is a measure of an overall spectral quality. Since a typical NMR signal consists of tens of thousands of data points, each bin can accommodate hundreds of data points required for an effective Benford analysis. In the case of a two-dimensional dataset, we may consider rectangular bins and shift them systematically all over the dataset, to obtain a 2D-SBA

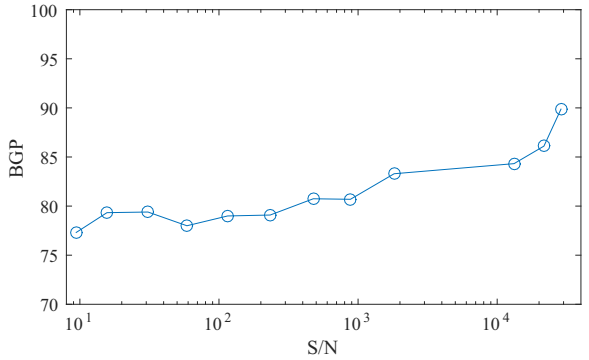

Figure 3. BGP versus signal to noise ratio for ^{1}H spectrum of water.