Astronomy and Computing 23 (2018) 151–165

Contents lists available at ScienceDirect

Astronomy and Computing

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

Full length article

Statistical detection of patterns in unidimensional distributions by continuous wavelet transforms

R.V. Baluev [*]

Central Astronomical Observatory at Pulkovo of the Russian Academy of Sciences, Pulkovskoje sh. 65/1, Saint Petersburg 196140, Russia
Saint Petersburg State University, Faculty of Mathematics and Mechanics, Universitetskij pr. 28, Petrodvorets, Saint Petersburg 198504, Russia

ARTICLE INFO

Article history:
Received 17 May 2017
Accepted 24 April 2018
Available online 2 May 2018

Keywords:
Methods: Data analysis
Methods: Statistical
Astronomical data bases: Miscellaneous
Planetary systems
Stars: Statistics
Galaxies: Statistics

ABSTRACT

Objective detection of specific patterns in statistical distributions, like groupings or gaps or abrupt transitions between different subsets, is a task with a rich range of applications in astronomy: Milky Way stellar population analysis, investigations of the exoplanets diversity, Solar System minor bodies statistics, extragalactic studies, etc. We adapt the powerful technique of the wavelet transforms to this generalized task, making a strong emphasis on the assessment of the patterns detection significance. Among other things, our method also involves optimal minimum-noise wavelets and minimum-noise reconstruction of the distribution density function. Based on this development, we construct a self-closed algorithmic pipeline aimed to process statistical samples. It is currently applicable to single-dimensional distributions only, but it is flexible enough to undergo further generalizations and development.

1. Introduction

Nowadays, the wavelet analysis technique is frequently used in various fields of astronomy. It proved a powerful tool in the time-series analysis, in particular to trace the time evolution of quasi-periodic variations (Foster, 1996; Vityazev, 2001). So far, the time-series analysis remains the major application domain of the wavelet transform, and most of the wavelet methodology and results are tied to this field. However, there are other branches where this technique appeared promising, in particular in the analysis of statistical distributions. In multiple astronomical applications we deal with statistical samples and distributions of various objects. For example, the last 20 years brought up a rich diversity of exoplanetary systems, and investigating their distributions gives a tremendous amount of information about the process of planet formation, their migration and dynamical evolution (Cumming, 2010). Other possible applications include the analysis of Milky Way stellar population that becomes more important with the emerging GAIA data (Brown et al., 2016), and the statistical analysis of minor bodies distributions in Solar System.

We emphasize that we formulate our goal here as 'patterns detection' rather than 'density estimation'. The latter would literally mean to estimate the probability density function (p.d.f.) of a sample, but instead we aim to detect easily-interpretable structures and shapes in this p.d.f., like e.g. clusters of objects, or paucities, or quick gradients. Such inhomogeneities often carry hidden knowledge about physical processes that objects of the sample underwent. In such a way, our task becomes related to data mining techniques and cluster analysis.

First attempts to apply the wavelet transform technique to reveal clumps in stellar distributions date back to 1990s (Chereul et al., 1998; Skuljan et al., 1999), and Romeo et al. (2003, 2004) suggested the use of wavelets for denoising results of N-body simulations. Nowadays, wavelet transforms are quite routinely used to analyse CMB data from WMAP (McEwen et al., 2004, 2017). The very idea of 'wavelets for statistics' is not novel too (Fadda et al., 1998; Abramovich et al., 2000).

When applied to these tasks, wavelets allow to objectivize the terms like the 'detail' or 'structural pattern' in a distribution, and easily formalize the task of 'patterns detection'. However, there are several crucial issues in this technique that either remain unresolved or solutions available in the literature look unsatisfactory and sometimes even flawed. In particular, the following matters raise questions.

1. Applying discrete wavelet transforms (DWTs) in this task seems unnatural. A major argument in favour of the DWT in 1990s might be to improve the computing performance. Nowadays, computing capabilities do not limit the practical use of continuous wavelet transforms (CWTs). Besides, the CWT mathematics is easier in many aspects.
2. Most if not all authors perform preliminary binning of the sample, or another kind of smoothing, before they apply a wavelet transform. It is an unnecessary and possibly even

[*] Correspondence to: Central Astronomical Observatory at Pulkovo of the Russian Academy of Sciences, Pulkovskoje sh. 65/1, Saint Petersburg 196140, Russia.
E-mail address: r.baluev@spbu.ru.

https://doi.org/10.1016/j.ascom.2018.04.004
2213-1337/© 2018 Elsevier B.V. All rights reserved.

harmful step. Small-scale structures are averaged out by the binning, and wavelets cannot 'see' them after that.

3. There is a big issue with correct determination of the statistical significance at the noise thresholding stage (discussed below).
4. It was not verified, whether the classic and typically used wavelets are indeed suitable in this task. Perhaps a systematic search is necessary, among wavelets of different shapes and utilizing some objective criteria of optimality.

A crucial problem of any statistical analysis is how to justify the statistical significance of the results. Determination of the statistical significance was always recognized as an important issue in this task. Unfortunately, approaches developed for the wavelet analysis of time series are not applicable for distribution analysis. Nonetheless, several schemes are available in the literature of how to do significance testing of the wavelet coefficients derived from a statistical sample (e.g. Skuljan et al., 1999; Fadda et al., 1998). But all these works share one common flaw: they define the significance of an *individual* wavelet coefficient, but test in turn *multiple* coefficients at once. In practice it leads to a dramatic increase of the false detections rate above the predicted level.

Consider that we perform N_t independent significance tests on the same sample, and every individual test is tuned to have a small enough false alarms probability (or p-value) β. The total number of false alarms is then $\sim \beta N_t$, and it may become unpredictably large because of large N_t. Let N_d be the number of significant ('detected') wavelet coefficients that passed the test. Then the fraction of false alarms *among the detected coefficients* is $\beta_{\mathrm{rel}} \sim \beta N_t/N_d$. Usually $\beta_{\mathrm{rel}} \gg \beta$, so the relative fraction of false alarms becomes much larger than the requested 'false alarm probability', paradoxically compromising the latter term. In other words, the false alarm probability is in fact misapplied in this task. In practice it may easily appear that *the majority* of the wavelet coefficients that formally passed their individual significance tests, appear in turn just noisy fluctuations.

In applications, the attention is paid to every detected detail of the distribution. Each false-detected wavelet coefficient trails a false 'pattern' in the recovered distribution. We guess that a researcher would expect that *all* structures that were claimed significant by the analysis algorithm, are significant indeed. So our intention is to narrow the 'false detection' term from 'an individual wavelet coefficient was wrongly claimed significant' to a more stringent 'at least one of many wavelet coefficients was wrongly claimed significant'. This triggers an effect generally similar to the one known as the 'bandwidth penaly' in the periodogram analysis of time series (Horne and Baliunas, 1986; Schwarzenberg-Czerny, 1998; Baluev, 2008).

In addition to what said above, only Monte Carlo simulations can currently be used to calculate the necessary p-values when testing the CWT significance. But numerical simulations are obviously inefficient, because they are very CPU-expensive and lack the generality. Some basic initial work on this problem was made in Baluyev (2005). In this paper we treat analytically the above-mentioned issues, and present the entire analysis pipeline.

2. Overview of the paper

In the literature there is a deficit of research dedicated to the task stated above, and there is a diversity of lesser sub-problems yet to be solved. Although many useful partial results are available, the very formalism of this task is still under construction, so there is no complete and self-consistent theory that we could use here 'as is'. Below we consider the following issues:

1. Adaptation of the CWT technique to statistical samples and distributions (Section 3);
2. Characterization of the noise that appears in the wavelet transform and construction of the signal detection criterion (Section 4);
3. Control of the non-Gaussian noise in the CWT that appears due to the small-number statistics and limits applicability of the entire technique (Section 4);
4. Search of optimal wavelets that improve the efficiency of the analysis (Section 5);
5. Optimal reconstruction of the distribution function itself from its CWT after noise thresholding (Section 6);
6. Numerical simulations and tests aimed to verify our theoretic results and constructions, demonstrate the main issues of the technique, and determine limits of its practical applicability (Section 7).

Finally, in Section 8 we provide a brief summary of our wavelet analysis algorithm.

3. Wavelet transforms

3.1. Basic definitions and formulae

We adopt the classic definition of the CWT from Grossman and Morlet (1984) with only a minor modification in scaling:

$$Y(a,b) = \int_{-\infty}^{+\infty} f(x)\psi\left(\frac{x-b}{a}\right) dx. \tag{1}$$

Here, $f(x)$ is an input function of the CWT. In this paper, it is meant to be a p.d.f. The kernel $\psi(t)$ is meant to be a *wavelet*. The latter term does not have a stable and strict definition, but at least ψ must be well localized together with its Fourier transform $\hat{\psi}$. Classic definitions also contain normalization factors in (1), typically $1/\sqrt{a}$, which we discard here.

The integral transform (1) is similar to a convolution, but contains two parameters: the scale a and the shift b. Contrary to the usual convolution, the CWT is easily invertible. Multiple inversion formulae are available, in particular based on Liu et al. (2015) we can write:

$$f(x) = \frac{1}{C_{\psi\gamma}} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} Y(a,b)\gamma\left(\frac{x-b}{a}\right)\frac{da db}{|a|^3},$$
$$C_{\psi\gamma} = \int_{-\infty}^{+\infty} \hat{\gamma}(\omega)\hat{\psi}^*(\omega)\frac{d\omega}{|\omega|}. \tag{2}$$

Here, the choice of the reconstruction kernel $\gamma(t)$ is rather arbitrary: it is mainly restricted by the mutual admissibility condition $0 < |C_{\psi\gamma}| < +\infty$. One of the most famous inversion formulae (Grossman and Morlet, 1984; Vityazev, 2001) contains $\gamma = \psi$, and it is similar to the original CWT (1). It requires that the wavelet must satisfy the classic admissibility condition $0 < C_{\psi\psi} < +\infty$, implying in particular that $\hat{\psi}(0) = 0$, and hence $\psi(t)$ must integrate to zero. This special case can be viewed as an orthogonal projection in the Hilbert space of Y, while other γ correspond to *oblique* projections.

The generalized inversion formulae (2) can be verified by applying the Fourier transform to it.

An alternative definition of the CWT can be written down as

$$\Upsilon(\kappa,s) = \int_{-\infty}^{+\infty} f(x)\psi(\kappa x + s)dx, \tag{3}$$

where $\kappa = 1/a$ is a wavenumber-like parameter, while $s = -b/a$ is a phase-like parameter. In terms of κ and s, the inversion formula (2) attains the following shape:

$$f(x) = \frac{1}{C_{\psi\gamma}} \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} \Upsilon(\kappa,s)\gamma(\kappa x + s)d\kappa ds. \tag{4}$$