ADVANCES IN DATA ANALYSIS AND CLASSIFICATION, 18 January (2018)

DOI: http://doi.org/10.1007/s11634-018-0309-2

New distance measures for classifying X-ray astronomy data into stellar classes*

Amparo Baíllo[a], Javier Cárcamo[a] and Konstantin Getman[b]

[a] Departamento de Matemáticas, Universidad Autónoma de Madrid, 28049 Madrid (Spain)

[b] Department of Astronomy and Astrophysics, Pennsylvania State University, University Park PA 16802-6305 (U.S.A.)

Abstract

The classification of the X-ray sources into classes (such as extragalactic sources, background stars, ...) is an essential task in astronomy. Typically, one of the classes corresponds to extragalactic radiation, whose photon emission behaviour is well characterized by a homogeneous Poisson process. We propose to use normalized versions of the Wasserstein and Zolotarev distances to quantify the deviation of the distribution of photon interarrival times from the exponential class. Our main motivation is the analysis of a massive dataset from X-ray astronomy obtained by the Chandra Orion Ultradeep Project (COUP). This project yielded a large catalog of 1616 X-ray cosmic sources in the Orion Nebula region, with their series of photon arrival times and associated energies. We consider the plug-in estimators of these metrics, determine their asymptotic distributions, and illustrate their finite-sample performance with a Monte Carlo study. We estimate these metrics for each COUP source from three different classes. We conclude that our proposal provides a striking amount of information on the nature of the photon emitting sources. Further, these variables have the ability to identify X-ray sources wrongly catalogued before. As an appealing conclusion, we show that some sources, previously classified as extragalactic emissions, have a much higher probability of being young stars in Orion Nebula. *Keywords:* Classification X-ray astronomy Wasserstein distance Zolotarev

metric Photon interarrival time Exponential distribution

Keywords: Classification; X-ray astronomy; Wasserstein distance; Zolotarev metric; photon interarrival time; exponential distribution.

MSC: Primary 60K35; secondary 62G20, 62N05.

*Research by A.B. and J.C. was supported by the Spanish MEyC grants MTM2013-44045-P and MTM2016-78751-P. K.G. acknowledges the support from the Chandra ACIS Team contract SV4-74018 (G. Garmire & L. Townsley, PIs), issued by the Chandra X-ray Center, which is operated by the Smithsonian Astrophysical Observatory on behalf of NASA under contract NAS8-03060.

1 Introduction and motivation

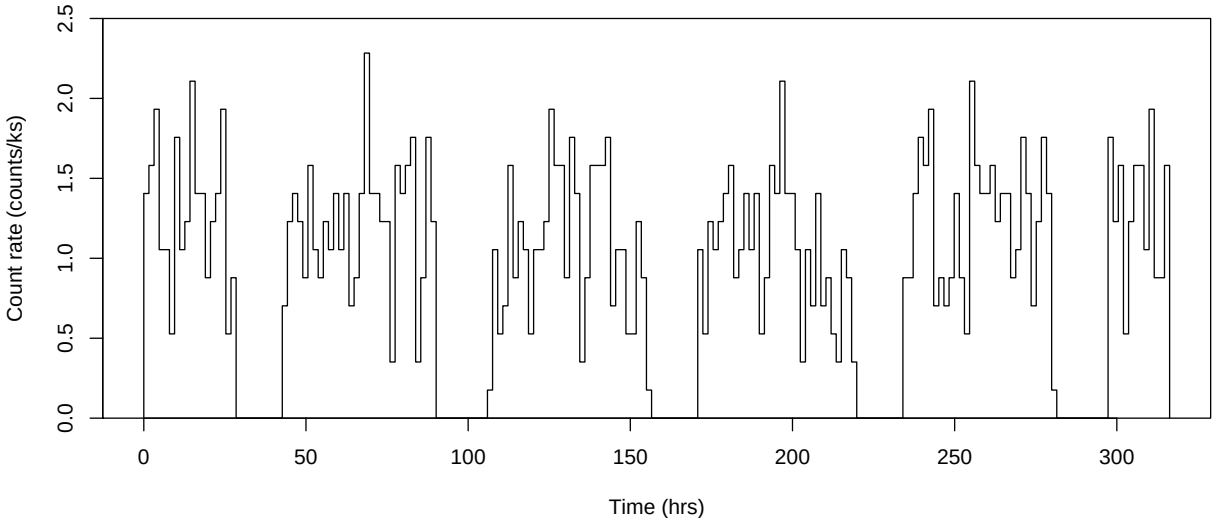
An important initial step in the analysis of stellar populations is the classification of samples into different classes of sources (see [5]). The definition of the classes (foreground stars, background stars, different types of pre-main-sequence stars, etc.) depends on the research project, but it is always of interest to identify extragalactic sources (see [5]; [9]). Frequently, the allocation has a degree of uncertainty, to the extent that some of the astronomical sources might remain unclassified (see [11]) or even wrongly catalogued.

X-ray astronomy deals with the detection and observation of astrophysical objects by means of the properties of their X-ray emissions. There are many astronomical sources of X-rays, such as galaxy clusters, black holes or different types of stars. In X-ray astronomy, classification of the data (that is, the X-ray sources) is accomplished using all the information provided by source features such as its location and X-ray and infrared properties (see [5]). As X-radiation is blocked by the atmosphere of Earth, cosmic X-ray emissions can only be detected by space telescopes.

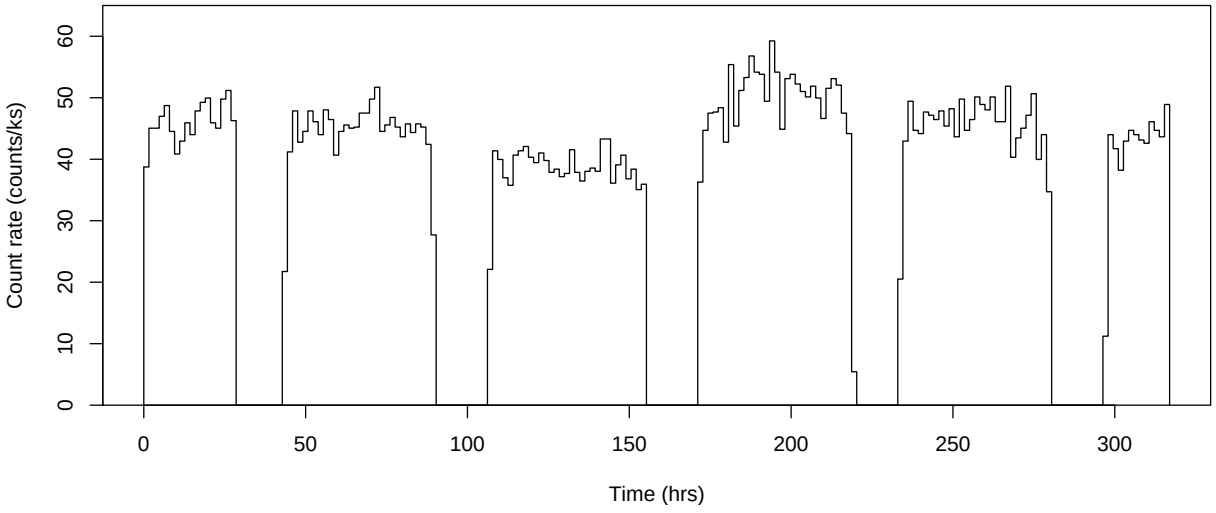
This article is motivated by a real dataset obtained as a result of Chandra Orion Ultradeep Project (COUP). It was fulfilled with one of the "Great Observatories" of NASA, the Chandra X-ray space telescope. Chandra was designed to observe X-ray emissions from high-energy regions of the space such as supernovas, black holes or star clusters as the Orion Nebula.

In this work, we focus on a massive collection of X-ray astronomical sources derived from a 2003 exposure of Chandra to the Orion Nebula region ([12]). For each of the sources captured by Chandra, the photon arrival times and associated energies were collected during a nearly continuous observation period of almost 10 days. The majority of these X-ray sources have been classified into one of three groups: lightly-obscured and heavily-obscured low-mass young stars; and extragalactic sources. The X-ray classification of young stellar objects in star forming regions is in general a complicated task, where numerous source properties are used as features.

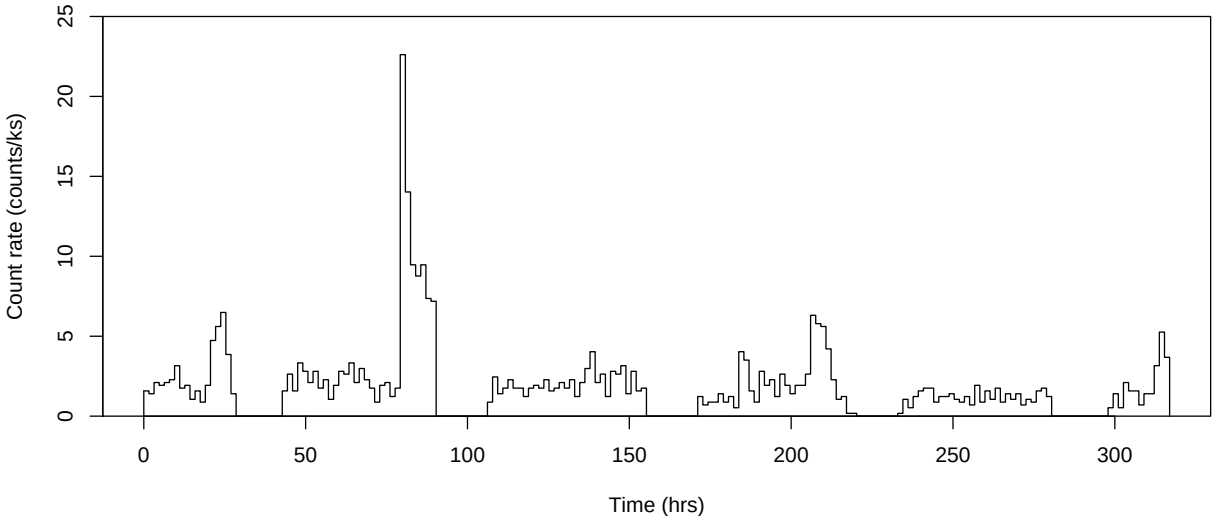
A very informative fact employed in this stellar classification is that, on the one hand, extragalactic radiation usually has a constant photon emission rate. This can be illustrated via the light curve of the astronomical source, a graph depicting its brightness (measured, e.g., by the photon count rate) over the course of the observation period. The light curves of COUP sources 111 (Figure 1 (a)) and 1304 (Figure 1 (b)), classified as extragalactic in [11] are examples of a constant photon arrival rate. In this case, the point process constituted by the photon arrival times is well-modeled by a homogeneous Poisson process. Thus, photon interarrival times from an extragalactic source should be close to an exponential distribution. On the other hand, young stars usually exhibit high-amplitude rapid variability ([29]) and their photon arrivals are generally affected by flares so the corresponding interarrival distribution might deviate from the exponential one. As an example, Figure 1 (c) displays the light curve of COUP source 89, corresponding to a young star, where we can see a large flare at about 80 hours from the start of the observation period. In the astrophysical literature, there are different proposals to quantify photon emission variability in stellar X-ray sources (see [29]).



(a)



(b)



(c)

Figure 1: Light curves for COUP sources (a) 111 (extragalactic radiation); (b) 1304 (extragalactic radiation); (c) 89 (lightly obscured PMS star).

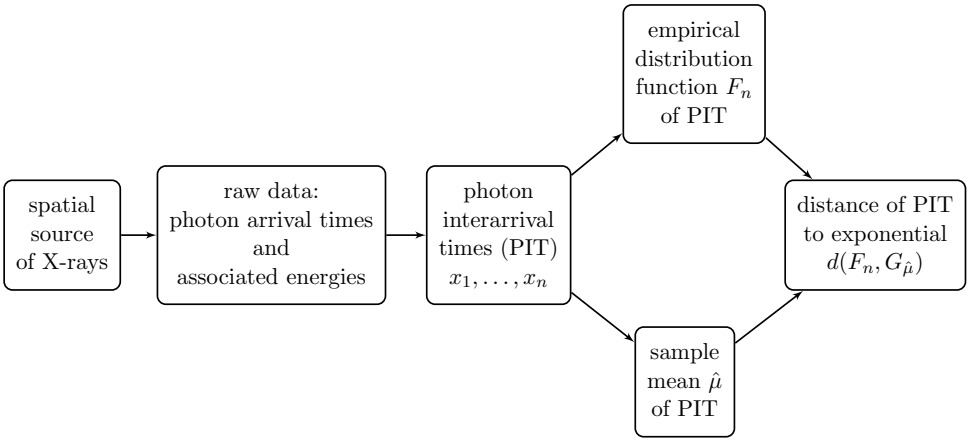


Figure 2: Data analysis pipeline for each single cosmic X-ray source (input): the raw data is the series of photon arrival times and their corresponding energies. PIT are computed as differences between consecutive arrival times. The sample mean and the empirical distribution function of PIT produce the output (the empirical distance of the PIT distribution to the exponential one).

Here, we propose a new statistical methodology to quantify the deviation of a random variable (namely, the photon interarrival time or PIT) from the exponential class. The final aim is to generate a new input variable for the discriminant procedure distinguishing among the source classes, and particulary extragalactic ones (see the data analysis pipeline in Figure 2). We consider that a large estimated distance of the PIT to the exponential class is an evidence that the corresponding source is not extragalactic. Specifically, we use a normalized version of the so-called Wasserstein and Zolotarev ζ_2 metrics, between the photon interarrival times of each X-ray source and the exponential distribution. As mentioned in [21, Section 15], the Zolotarev ζ_2-metric is appropriate when dealing with exponential variables. Further, [22] argue that Wasserstein and Zolotarev distances are more sensitive to extreme values than other probability metrics such as the usual Kolmogorov distance. In general, it is often desirable to take into account extreme events to compare distributions. This is specially relevant with data coming from astrophysical studies (see [8]). We demonstrate that these distances can be used as informative variables that detect groups or similitudes among the X-ray sources and help identify possible outliers within a group. In this work the term "outlier" refers to a source whose distance to the exponential distribution is substantially different from the distances of other members of the same class. In fact, in the final analysis of the COUP data, we show that some of the outlying COUP sources, initially classified as extragalactic radiation, could actually be young stars in the Orion Nebula region.

The ideas in this paper are also potentially useful in other biological or physical problems in which deviations from the exponential model need to be detected and quantified. Alternatively, the proposed methodology allows assessing whether a homogeneous Poisson process achieves a good approximation of an observed phenomenon.

This paper is structured as follows. In the next section, we describe in detail the COUP dataset. In Section 3, we introduce the Wasserstein and Zolotarev distances and normalized versions of them. We also consider the plug-in estimators of these metrics to be used in practice and determine their asymptotic distributions. In Section 4, we carry out a simulation study to assess the practical performance of our proposal with finite samples. The COUP dataset is analyzed in depth in Section 5. Finally, the main conclusions are collected in Section 6. In the Appendix, we include the proofs of the results stated in Section 3 and other technical details.

2 Chandra Orion Ultradeep Project Dataset

Among other things, the COUP analyzes X-flaring in pre-main-sequence (PMS) stars, members of the Orion Nebula region that is composed of the rich revealed Orion Nebula Cluster (ONC) and the filamentary molecular cloud called Orion Molecular Cloud 1 (OMC-1). A PMS star is a premature star that has acquired all of its mass from its natal envelope of interstellar dust and gas and contracts until it starts hydrogen burning (see, e.g., [23]). These young stars have intense magnetic fields, detected through their X-ray emissions, where plasma, confined in magnetic loops, is heated to X-ray emitting temperatures. Detection of these high-energy emissions is only possible by space observatories, such as the Chandra X-ray Observatory (see `http://chandra.si.edu/`). The nearest rich and concentrated collection of PMS stars is in the ONC/OMC-1 star forming region.

In January 2003, the Chandra X-ray Observatory focused its Advanced CCD Imaging Spectrometer on the ONC for a period of 13.2 days obtaining the deepest X-ray observation ever taken of any star cluster (see [12]). This observation period was only interrupted by the five passages of the Chandra spacecraft through the Van Allen radiation belts. Graphically, this is reflected in the five gaps appearing in the light curves of Figure 1. The results were an almost continuous observation of the photon arrival times and associated energies for 1616 X-ray sources. COUP sources were compared with source positions from previously existing catalogs, with the aim of physically associating the COUP sources with already identified stars (whenever this was possible). The majority of these COUP sources has been classified into one of three groups (see [9]):

- *Lightly-obscured* PMS sources: This class is constituted by 835 cool low-mass PMS stars that are likely located in the ONC cluster. The term "lightly obscured" means that the star X-ray emission is less absorbed by the material in the interstellar medium.
- *Heavily-obscured* PMS sources: This class corresponds to 559 low-mass PMS stellar objects that are likely still embedded in the nascent OMC-1 cloud.
- *Nonmembers*: This group contains over 200 probable nonmembers of the Orion Nebula star forming region. A large part are likely extragalactic sources and a few are foreground stars or very faint sources without counterparts. For our analysis, in this group we only consider the extragalactic X-ray emissions.

The original classification of these 1594 COUP sources is detailed in [11], who provide an estimated probability of membership to the Orion cloud for each source. Thus, this

classification has a degree of uncertainty. The aim of this work is to introduce new distance measures that contain relevant information for classifying X-ray astronomy data into stellar classes. Specifically, we compute some probability metrics of the PIT to the exponential class, as they exhibit different distributions depending on the nature of the photon emission sources. These distances can be incorporated in any classification rule to reduce the classification error.

3 Quantifying the discrepancy from extragalactic radiation

As mentioned in the introduction, PIT resulting from extragalactic radiation are usually well-modeled by exponential distributions while, for the PMS stars classes, PIT distributions might deviate from the exponential one. Our primary objective is to compare different sources by computing some normalized probability metrics between their corresponding PIT and the nearest exponential variable, and to highlight the classification power of these features. This idea has been tackled before: for instance, in X-ray astronomy the usual Kolmogorov distance to the exponential class is often used in data classification. However, depending on the problem at hand, other distances could be more appropriate, e.g., to take into account extreme values or to highlight a specific part of the distribution. Here, we analyze in depth the performance of the Wasserstein and Zolotarev distances, more sensitive to the behaviour in the tail of the distribution than others such as the usual Kolmogorov and Cramér-von Mises metrics. In particular, a byproduct of the results in this section is the possibility of testing for exponentiality (see the end of Subsection 3.2), for which there are numerous proposals in the literature (see the reviews by [2] and [16]).

3.1 The choice of the metrics

A central question in the problem under consideration is the choice of a suitable metric to measure how far the probability distribution of the positive random variable X of interest (the PIT), with expectation $\mu > 0$, is from the exponential variable Y_μ with the same mean. In this work, we focus on the family of integral probability metrics

$$d_r(X, Y_\mu) = \sup \{|\mathrm{E}f(X) - \mathrm{E}f(Y_\mu)| : f \in \mathcal{F}_r\}, \quad r \in \mathbb{N}, \tag{1}$$

where $\mathcal{F}_r$ is the class of real-valued functions f on $\mathbb{R}$ having r-th derivative $f^{(r)}$ a.e. and such that $|f^{(r)}| \leq 1$ a.e. Observe that $d_r(X, Y_\mu)$ is the maximum error in the expected value within the class of smooth functions $\mathcal{F}_r$ due to the approximation of X by Y_μ. For notational convenience, if F and G_μ are the distribution functions of X and Y_μ, respectively, we indistinctly use $d_r(F, G_\mu)$ or $d_r(X, Y_\mu)$. Note also that $G_\mu(x) = 1 - \exp(-x/\mu)$, for $x \geq 0$.

The case $r = 1$ in (1) has special relevance. By the Kantorovich–Rubinstein theorem, we see that $d_1 \equiv \omega$ is the famous L^1-*Wasserstein distance*. For $r \geq 2$, $d_r \equiv \zeta_r$ is the *Zolotarev metric* of order r. For a general reference on these distances we refer to [21].

It can be proved that, when $d_r(X, Y_\mu) < \infty$, the moments of X and Y_μ coincide up to order $r - 1$. As a consequence, in general, when $r \geq 3$ and X is *not* exponential, we have that $d_r(X, Y_\mu) = \infty$. This follows from the fact that, for many distributions, equalities $\mathrm{E}X = \mu$ and $\mathrm{E}X^2 = 2\,\mu^2$ are too restrictive and they actually imply that X is exponential. For

instance, this happens for the variables with the HNBUE or HNWUE property, two large families of random variables that include all the usual ageing and anti-ageing classes of distributions as it follows from results on stochastic equality under convex domination (see, e.g., [25, Theorem 3.A.42, p. 133]). Therefore, only the cases $r = 1$ and $r = 2$ make sense for the discussed problem, that is, the Wasserstein distance) $d_1 \equiv \omega$ and the Zolotarev metric $d_2 \equiv \zeta_2$. These two metrics have easier-to-handle dual integral representations (see [22]), given by

$$\omega(F, G_\mu) = \int_0^\infty |F(t) - G_\mu(t)| \, \mathrm{d}t, \tag{2}$$

$$\zeta_2(F, G_\mu) = \int_0^\infty \left| \int_t^\infty (F(x) - G_\mu(x)) \, \mathrm{d}x \right| \, \mathrm{d}t. \tag{3}$$

The metrics ω and ζ_2 have practical advantages for the problem at hand. First, as argued in [22, p. 15], they are more sensitive to the differences in the probabilities corresponding to extreme values than other common probability metrics such as the *Kolmogorov distance*, $\kappa(F, G) = \sup_x |F(x) - G(x)|$. Since the difference $|F(x) - G(x)|$ converges to zero as x tends to $+\infty$ or $-\infty$, the contribution of the terms corresponding to extreme events is usually small. As a consequence, the differences in the tail behavior of X and Y will only be reflected in $\kappa(F, G)$ to a relatively small extent. However, representations (2) and (3) show that extreme values have more weight in ω and ζ_2, as integrals of tail probabilities appear in these distances. Additionally, Zolotarev ζ_2-metric is considered as the "natural metric" when dealing with the exponential class (see [21, p. 340]).

It becomes soon apparent that these two metrics have the important drawback of not being scale-independent, an essential requirement to compare X-ray sources with very different photon emission rates (the average of photon emissions in each time interval). Even for sources of the same nature, the distances ω and ζ_2 to the exponential distribution could be extremely different. This is clearly reflected in Figure 1, where COUP 111 and 1304, both classified as extragalactic, would not be comparable without a suitable normalization.

To overcome this problem, next we introduce dimensionless versions of the ω and ζ_2 distances. Let us observe that the Wasserstein ω and Zolotarev ζ_2 metrics are homogeneous of order 1 and 2, respectively. We define the *normalized Wasserstein* and *normalized Zolotarev* metrics as

$$\bar{\omega}(X, Y_\mu) = \omega(X/\mu, Y_\mu/\mu) = \frac{1}{\mu}\omega(X, Y_\mu)$$

and

$$\bar{\zeta}_2(X, Y_\mu) = \zeta_2(X/\mu, Y_\mu/\mu) = \frac{1}{\mu^2}\zeta_2(X, Y_\mu).$$

These are the two probability distances that should be used in practice (instead of their unnormalized versions), as they are homogeneous of degree 0 and dimensionless.

3.2 Estimation and large sample behaviour

In practice, the distribution of the random variable X is usually unknown. Therefore, the distances $d(X, Y_\mu)$ (for $d = \omega, \zeta_2, \bar{\omega}$ and $\bar{\zeta}_2$) have to be estimated using a random sample

$X_1, \ldots, X_n$ from X. We propose to use the plug-in estimators obtained by replacing the true distribution F of X by the empirical distribution F_n of the sample $X_1, \ldots, X_n$, that is, $F_n(t) = n^{-1} \sum_{i=1}^n I_{\{X_i \le t\}}$, $n \in \mathbb{N}$, $t \ge 0$, where I_A stands for the indicator function of the set A. Thus, we estimate $d(F, G_\mu)$ by $d(F_n, G_{\hat{\mu}})$, where $\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$ is the sample mean (and the maximum likelihood estimator of the rate parameter μ of an exponential model).

Analyzing the asymptotic behavior of the empirical distances $d(F_n, G_{\hat{\mu}})$ is an important issue to understand its performance and accuracy in practice, for instance, to assess whether an exponential model provides a reasonably good approximation of X. Besides, the asymptotic probability distribution potentially allows performing inference on $d(X, Y_\mu)$. For the considered distances, we note that

$$d(F_n, G_{\hat{\mu}}) = d(F, G_\mu) + \frac{1}{\sqrt{n}} \delta_n(d, F),$$

where $\delta_n(d, F)$ is the standardized version of the estimated distances

$$\delta_n(d, F) := \sqrt{n} \left(d(F_n, G_{\hat{\mu}}) - d(F, G_\mu) \right), \quad n \in \mathbb{N}. \tag{4}$$

Here we find conditions (as sharp as possible) on the random variable X so that $\delta_n(d, F)$ converges in distribution as $n \to \infty$, and determine its weak limit, $\delta_\infty(d, F)$. In this way, we obtain that

$$d(F_n, G_{\hat{\mu}}) = d(F, G_\mu) + O_{\mathrm{P}}(1/\sqrt{n}), \quad \text{as } n \to \infty.$$

Though the detailed proofs of the asymptotic distribution of $\delta_n(d, F)$ are collected in the Appendix, we describe here in broad strokes the main ideas behind them. First, we note that

$$\delta_n(d, F) = \rho_n(\mathbb{X}_{d,n}, g_d), \tag{5}$$

where $\rho_n : L^1 \times L^1 \to \mathbb{R}$ is the functional defined by

$$\rho_n(f, g) := \|f + \sqrt{n} g\|_1 - \sqrt{n} \|g\|_1, \quad \text{for } f,\, g \in L^1, \tag{6}$$

$\mathbb{X}_{d,n}$ are the stochastic processes given (for $t \ge 0$) by

$$\begin{aligned}
\mathbb{X}_{\omega,n}(t) &:= \sqrt{n} \left[(F_n(t) - G_{\hat{\mu}}(t)) - (F(t) - G_\mu(t)) \right], \\
\mathbb{X}_{\bar{\omega},n}(t) &:= \sqrt{n} \left[\frac{1}{\hat{\mu}} (F_n(t) - G_{\hat{\mu}}(t)) - \frac{1}{\mu} (F(t) - G_\mu(t)) \right], \\
\mathbb{X}_{\zeta_2,n}(t) &:= \sqrt{n} \left[\int_t^\infty (F_n(x) - G_{\hat{\mu}}(x)) \, \mathrm{d}x - \int_t^\infty (F(x) - G_\mu(x)) \, \mathrm{d}x \right], \\
\mathbb{X}_{\bar{\zeta}_2,n}(t) &:= \sqrt{n} \left[\frac{1}{\hat{\mu}^2} \int_t^\infty (F_n(x) - G_{\hat{\mu}}(x)) \, \mathrm{d}x - \frac{1}{\mu^2} \int_t^\infty (F(x) - G_\mu(x)) \, \mathrm{d}x \right],
\end{aligned} \tag{7}$$

and g_d are the (deterministic) functions defined by

$$g_\omega(t) := F(t) - G_\mu(t), \qquad g_{\bar{\omega}}(t) := \frac{1}{\mu} (F(t) - G_\mu(t)), \tag{8}$$

$$g_{\zeta_2}(t) := \int_t^\infty (F(x) - G_\mu(x)) \, \mathrm{d}x, \qquad g_{\bar{\zeta}_2}(t) := \frac{1}{\mu^2} \int_t^\infty (F(x) - G_\mu(x)) \, \mathrm{d}x. \tag{9}$$

From (5), we see that establishing the (weak) convergence in L^1 of the processes $\mathbb{X}_{d,n}$ in (7), combined with the continuity of the linking functional in (6), immediately translates into the convergence in distribution of $\delta_n(d, F)$.

Before stating the main results, we need to introduce some definitions and notation. In the sequel, $\mathbb{B}_F := \mathbb{B} \circ F$ is the *F-Brownian bridge*, where $\mathbb{B}$ is a standard Brownian bridge on $[0, 1]$, that is, $\mathbb{B}$ is a centered Gaussian process with covariance function $\gamma(s, t) = \min(s, t) - st$ and continuous paths, with probability 1.

We consider the *Lorentz spaces* of positive random variables defined by $\mathcal{L}^{2,1} := \{X : \Lambda_{2,1}(X) < \infty\}$ and $\mathcal{L}^{4,2} := \{X : \Lambda_{4,2}(X) < \infty\}$, where

$$\Lambda_{2,1}(X) := \int_0^\infty \sqrt{\mathrm{P}(X > t)}\,\mathrm{d}t \quad \text{and} \quad \Lambda_{4,2}(X) := \int_0^\infty t\sqrt{\mathrm{P}(X > t)}\,\mathrm{d}t$$

(see [20, p. 279]). Conditions $\Lambda_{2,1}(X) < \infty$ and $\Lambda_{4,2}(X) < \infty$ are slightly stronger than $\mathrm{E}X^2 < \infty$ and $\mathrm{E}X^4 < \infty$, respectively (see [15]). Finally, $\mathcal{L}^p := \{X : \mathrm{E}X^p < \infty\}$ $(p > 0)$ is the usual space of (positive) random variables with finite p-th moment.

The following two theorems characterize the asymptotic behavior of $\mathbb{X}_{d,n}$ in L^1. The results are sharp in the sense that we obtain the exact integrability condition on X so that the processes converge in distribution in L^1 as $n \to \infty$. The symbol "$\xrightarrow{L^1}_{\mathrm{w}}$" stands for the weak convergence of a sequence of random processes in the space L^1 as $n \to \infty$ (see the Appendix for the precise definition).

Theorem 1. *Let X be a positive random variable with expectation $\mu > 0$. If $X \in \mathcal{L}^{4/3}$, the following assertions are equivalent:*

(a) $X \in \mathcal{L}^{2,1}$.

(b) $\mathbb{X}_{\omega,n} \xrightarrow{L^1}_{\mathrm{w}} \mathbb{X}_{\omega,F}$, *where* $\mathbb{X}_{\omega,F}$ *is a centered Gaussian process given by*

$$\mathbb{X}_{\omega,F}(t) := \mathbb{B}_F(t) - \frac{t}{\mu^2}e^{-t/\mu}\int_0^\infty \mathbb{B}_F(s)\,\mathrm{d}s, \quad t \geq 0. \tag{10}$$

(c) $\mathbb{X}_{\bar{\omega},n} \xrightarrow{L^1}_{\mathrm{w}} \mathbb{X}_{\bar{\omega},F}$, *where* $\mathbb{X}_{\bar{\omega},F}$ *is a centered Gaussian process given by*

$$\mathbb{X}_{\bar{\omega},F}(t) := \frac{1}{\mu}\left[\mathbb{B}_F(t) + \left(g_{\bar{\omega}}(t) - \frac{t}{\mu^2}e^{-t/\mu}\right)\int_0^\infty \mathbb{B}_F(s)\,\mathrm{d}s\right], \quad t \geq 0,$$

and the function $g_{\bar{\omega}}$ is defined in (8).

Theorem 2. *Let X be a positive random variable with expectation $\mu > 0$. The following assertions are equivalent:*

(a) $X \in \mathcal{L}^{4,2}$.

(b) $\mathbb{X}_{\zeta_2,n} \xrightarrow{L^1}_{\mathrm{w}} \mathbb{X}_{\zeta_2,F}$, *where* $\mathbb{X}_{\zeta_2,F}$ *is a centered Gaussian process given by*

$$\mathbb{X}_{\zeta_2,F}(t) := \int_t^\infty \mathbb{B}_F(s)\,\mathrm{d}s - \left(1 + \frac{t}{\mu}\right)e^{-t/\mu}\int_0^\infty \mathbb{B}_F(s)\,\mathrm{d}s, \quad t \geq 0.$$

(c) $\mathbb{X}_{\bar{\zeta}_2,n} \xrightarrow{L^1}_{\mathrm{w}} \mathbb{X}_{\bar{\zeta}_2,F}$, *where* $\mathbb{X}_{\bar{\zeta}_2,F}$ *is a centered Gaussian process given by*

$$\mathbb{X}_{\bar{\zeta}_2,F}(t) := \frac{1}{\mu^2}\left[\int_t^\infty \mathbb{B}_F(s)\,\mathrm{d}s + \left(2\,\mu\, g_{\bar{\zeta}_2}(t) - \left(1+\frac{t}{\mu}\right)e^{-t/\mu}\right)\int_0^\infty \mathbb{B}_F(s)\,\mathrm{d}s\right], \tag{11}$$

for $t \geq 0$, *and the function* $g_{\bar{\zeta}_2}$ *is defined in (9).*

Using (5) and Theorems 1 and 2, in the next theorem we derive the asymptotic distribution of $\delta_n(d,F)$. In the sequel "$\longrightarrow_{\mathrm{d}}$" stands for convergence in distribution as $n\to\infty$, $\mathrm{sgn}(\cdot)$ denotes the sign function and A^c is the complement of the set A.

Theorem 3. *Let* X *be a positive random variable with expectation* $\mu>0$. *For* $d=\omega$ *or* $d=\bar{\omega}$ *(respectively, for* $d=\zeta_2$ *or* $d=\bar{\zeta}_2$*), let us assume that* $X\in\mathcal{L}^{2,1}$ *(respectively,* $X\in\mathcal{L}^{4,2}$*). Then,* $\delta_n(d,F)\longrightarrow_{\mathrm{d}}\delta_\infty(d,F)$, *with*

$$\delta_\infty(d,F) := \int_{I(g_d)} |\mathbb{X}_{d,F}(t)|\,\mathrm{d}t + \int_{I(g_d)^c} \mathbb{X}_{d,F}(t)\,\mathrm{sgn}(g_d(t))\,\mathrm{d}t, \tag{12}$$

where the processes $\mathbb{X}_{d,F}$ *are defined in (10)-(11), the functions* g_d *are given in (8)-(9), and* $I(g_d) := \{t\geq 0 : g_d(t)=0\}$.

The next corollary, a direct consequence of Theorem 3, provides the asymptotic distribution of $\delta_n(d,F)$, when F is an exponential distribution function. In such a case, $d(F,G_\mu)=0$ and the estimators behave as a random quantity at the order of $O_{\mathrm{P}}(1/\sqrt{n})$. It is interesting to note that the limiting distribution of the normalized distances does not depend on the unknown mean of the exponential distribution.

Corollary 1. *For the processes* $\mathbb{X}_{d,F}$ *defined in (10)-(11), if* X *follows an exponential distribution with mean* μ, *then*

(a) $\sqrt{n}\,\omega(F_n,G_{\hat{\mu}}) \longrightarrow_{\mathrm{d}} \|\mathbb{X}_{\omega,G_\mu}\|_1 = \mu\,\|\mathbb{X}_{\omega,G_1}\|_1$;

(b) $\sqrt{n}\,\bar{\omega}(F_n,G_{\hat{\mu}}) \longrightarrow_{\mathrm{d}} \|\mathbb{X}_{\bar{\omega},G_1}\|_1 = \|\mathbb{X}_{\omega,G_1}\|_1$;

(c) $\sqrt{n}\,\zeta_2(F_n,G_{\hat{\mu}}) \longrightarrow_{\mathrm{d}} \|\mathbb{X}_{\zeta_2,G_\mu}\|_1 = \mu^2\,\|\mathbb{X}_{\zeta_2,G_1}\|_1$;

(d) $\sqrt{n}\,\bar{\zeta}_2(F_n,G_{\hat{\mu}}) \longrightarrow_{\mathrm{d}} \|\mathbb{X}_{\bar{\zeta}_2,G_1}\|_1 = \|\mathbb{X}_{\zeta_2,G_1}\|_1$.

The following corollary states that when X does not share any part of its distribution function with the exponential one, the limiting distribution $\delta_\infty(d,F)$ in (12) is actually normal, for all the considered distances.

Corollary 2. *Let us assume that the conditions of Theorem 3 hold and let us further assume that the set* $I(g_d)$ *(defined in Theorem 3) has zero Lebesque measure. Then,* $\delta_\infty(d,F)$ *defined in (12) has a zero mean normal distribution.*

The previous results could be useful to compute (asymptotic) confidence intervals for $d(F,G_\mu)$. This can be implemented via the following bootstrap procedure: as F is usually

unknown, we substitute F for F_n in the limiting processes obtained in Theorems 1 and 2 (equations (10)-(11)). Next, we simulate a large number of trajectories of the processes to obtain a Monte Carlo approximation of the asymptotic distribution $\delta_\infty(d, F)$ given in Theorem 3. Finally, we use the Monte Carlo sample quantiles to construct the desired interval. We also note that if we assume that the Lebesgue measure of the sets $I(g_d)$ is zero, then the procedure is simpler as Corollary 2 ensures that the limit distribution is a zero mean normal distribution. Hence, in such a case it is enough to estimate the asymptotic variance of the limit via Monte Carlo and use the quantiles of a normal distribution. This latter interval is called *standard normal interval* in [7, p. 168]. However, let us observe that asymptotic confidence intervals could be unprecise when the sample size is small. As the distribution of $\delta_n(d, F)$ for a fixed n could be extremely difficult to handle, an interesting alternative is to construct a bootstrap confidence interval of $d(F, G_\mu)$. In this situation, the *percentile interval* proposed in [7, p. 170] is a reasonable choice.

As $d(F, G_\mu) = 0$ is equivalent to saying that F is exponential, the considered distances can be additionally applied to goodness-of-fit tests for $\mathrm{H}_0 : F$ is exponential. As stated in Corollary 1, for $d = \bar{\omega}, \bar{\zeta}_2$, the asymptotic distribution of the test statistic $\sqrt{n}\, d(F_n, G_{\hat{\mu}})$ is completely determined under H_0. In practice, this result allows us to derive an asymptotic rejection region by Monte Carlo sampling from the asymptotic distribution. When the sample size n is small, a better alternative is to use a parametric bootstrap procedure by sampling from an exponential distribution with mean $\hat{\mu}$.

4 Simulations

The aim of this section is to analyze the finite-sample behaviour of the statistic $\delta_n(d, F)$ given in (4), in particular to compare it with its asymptotic distribution $\delta_\infty(d, F)$ obtained in Theorem 3. We only consider the normalized versions of the distances, $d = \bar{\omega}$ and $d = \bar{\zeta}_2$. In any case, the simulations results for the unnormalized metrics are similar and do not add relevant information.

To compute the normalized empirical distances $\bar{\omega}(F_n, G_{\hat{\mu}}) = \omega(F_n, G_{\hat{\mu}})/\hat{\mu}$ and $\bar{\zeta}_2(F_n, G_{\hat{\mu}}) = \zeta_2(F_n, G_{\hat{\mu}})/\hat{\mu}^2$, we have used the following equalities

$$\omega(F_n, G_{\hat{\mu}}) = X_{(1)} - \hat{\mu}\, G_{\hat{\mu}}(X_{(1)}) + \int_{X_{(1)}}^{X_{(n)}} |F_n(x) - G_{\hat{\mu}}(x)|\, \mathrm{d}x + \hat{\mu}\, e^{-X_{(n)}/\hat{\mu}} \tag{13}$$

and

$$\begin{aligned}\zeta_2(F_n, G_{\hat{\mu}}) &= 2 \int_0^\infty \left(\int_0^t (F_n(x) - G_{\hat{\mu}}(x))\, \mathrm{d}x \right)_+ \mathrm{d}t + \hat{\mu}^2 - \frac{a_2}{2} \\ &= 2 \int_{X_{(1)}}^{X_{(n)}} \left(-X_{(1)} + \hat{\mu}\, G_{\hat{\mu}}(X_{(1)}) + \int_{X_{(1)}}^t (F_n(x) - G_{\hat{\mu}}(x))\, \mathrm{d}x \right)_+ \mathrm{d}t + \hat{\mu}^2 - \frac{a_2}{2}, \end{aligned} \tag{14}$$

where $X_{(1)} \leq \cdots \leq X_{(n)}$ are the order statistics of the sample and $a_2 := \sum_{i=1}^n X_i^2/n$. Even though the integrals appearing in (13) and (14) can be expressed in terms of the order statistics, from a computational viewpoint it is more convenient to approximate them numerically: this was carried out by discretizing the integral on the equispaced grid $X_{(1)} + k\,\delta$ with $k = 0, 1, 2, \ldots, 20000$ and $\delta = (X_{(n)} - X_{(1)})/20000$.

The asymptotic distribution $\delta_\infty(d, F)$ can be approximately sampled by generating trajectories of the Brownian bridge $\mathbb{B}_F(t)$ on a bounded time interval $[0, T]$ and then approximating the integrals such as $\int_0^\infty \mathbb{B}_F(t)\,\mathrm{d}t$ by their discretization on $[0, T]$. In our work, we have chosen T as the smallest integer larger or equal to $F^{-1}(1 - \texttt{tol})$, where `tol` is a tolerance limit equal to 10^{-6}. The integral discretization was carried out with an equispaced grid on $[0, T]$ yielding 50000 subintervals.

Computing the normalized version, $d = \bar{\omega}$ or $d = \bar{\zeta}_2$, of the distances $d(F, G_\mu)$ is equivalent to computing the distance of the re-scaled variable X/μ to the exponential distribution with mean $\mu = 1$. As a consequence, in this Monte Carlo study the data-generating distributions have expectation $\mu = 1$. Specifically, we have considered:

- the exponential distribution with mean $\mu = 1$;
- the Weibull distribution with shape parameter $a > 0$ and scale parameter $\lambda = 1/\Gamma(1 + 1/a)$, with probability density $f(x) = a/\lambda\,(x/\lambda)^{a-1}\,e^{-(x/\lambda)^a}$, for $x > 0$;
- the gamma distribution with shape parameter $a > 0$ and scale parameter $\lambda = 1/a$, with density $f(x) = a^a/\Gamma(a)\,x^{a-1}\,e^{-ax}$, for $x > 0$.

For the Weibull distribution with mean 1 we have chosen the values $a = 0.9$ and $a = 1.1$ for the shape parameter. For the gamma distribution with mean 1, we have used shape parameters $a = 0.9$ and $a = 1.1$ (see Figure 3).

In the simulations we have generated 10000 samples of size $n = 100$, 500, 1000 and 5000 from each of these distributions. For each Monte Carlo sample, we have computed the statistic $\delta_n(d, F)$ for $d = \bar{\omega}$ and $d = \bar{\zeta}_2$. The programming language used in this work is R (`www.R-project.org`). The results of the simulations are summarized in Figures 4, 5, 6, 7 and 8. Each figure shows the evolution, as n increases, of the boxplots of these statistics $\delta_n(d, F)$ towards the boxplot of its asymptotic distribution $\delta_\infty(d, F)$ given in (12). This latter boxplot is also based on 10000 samples from the corresponding limit distribution.

On the one hand, we observe that, the finite-sample behavior of $\delta_n(d, F)$, both for $d = \bar{\omega}$ and for $d = \bar{\zeta}$, is stable for the exponential distribution. In particular, the quartiles and median of $\delta_n(d, F)$ are very similar for any n. For nonexponential distributions F, the boxplot of $\delta_n(d, F)$ resembles that of $\delta_\infty(d, F)$ for large sample sizes ($n \geq 1000$). On the other hand, we also observe that, the closer F is to the exponential distribution, the larger n has to be for the distribution of $\delta_n(d, F)$ to approach its limit. This is reasonable: when F is not exponential, but close to it and the set $I(g_d)$ has zero Lebesgue measure, the finite sample behavior of $\delta_n(d, F)$ is almost as if F were exponential, but the limit is actually Gaussian (see Corollary 2).

5 Analysis of the COUP data

As mentioned in the introduction, photon interarrival times resulting from extragalactic radiation are usually well-modeled by exponential distributions, while, for the classes of PMS stars, PIT distributions might deviate from the exponential one. Hence, it is natural to compare different sources by computing the distances between their corresponding PIT and an exponential variable with the same mean. In this section, we compute the

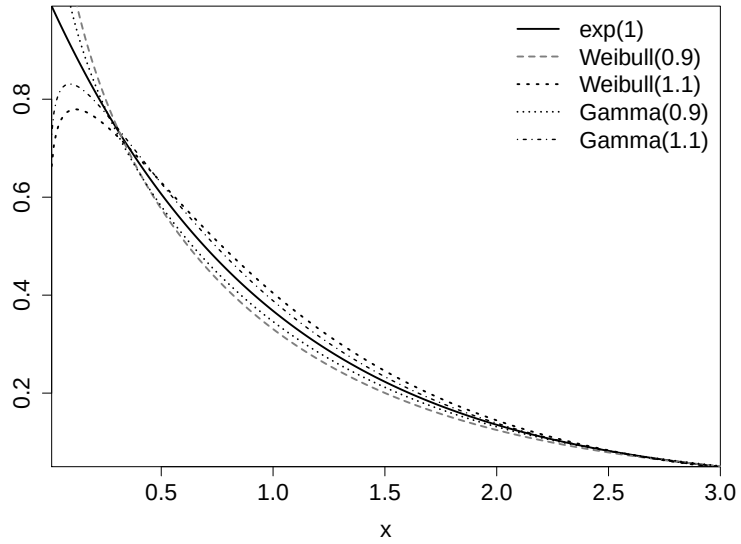
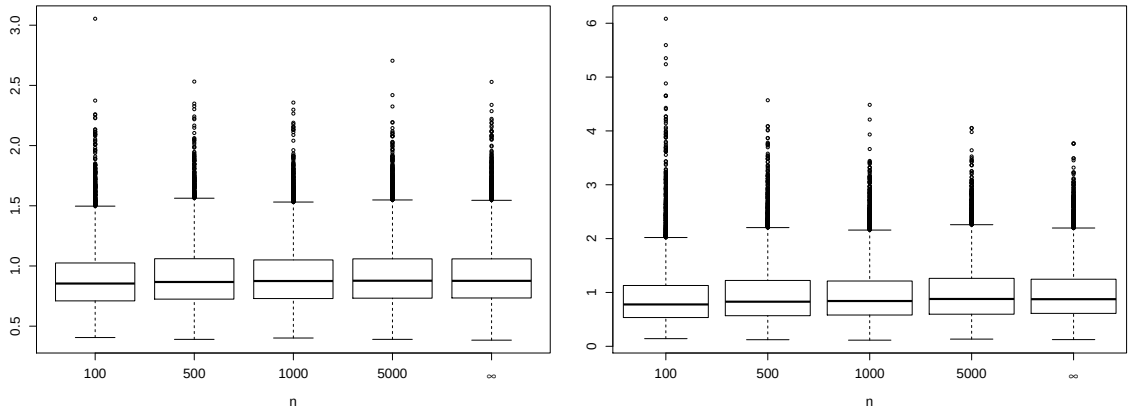
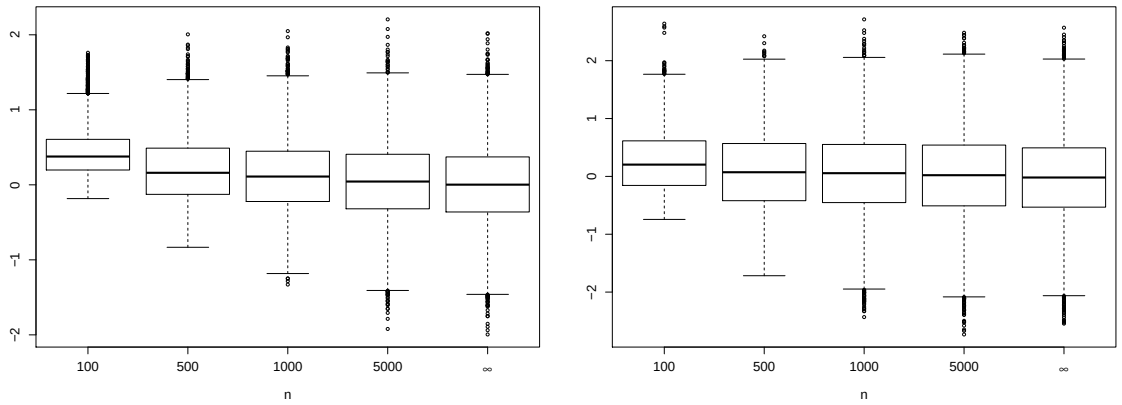


Figure 3: Probability densities in the simulation study of Section 4.



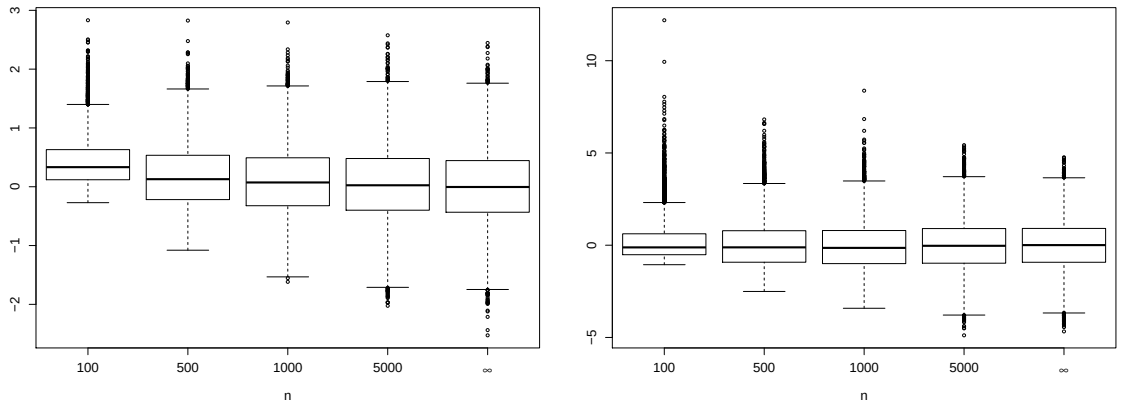
(a) (b)

Figure 4: Boxplots of $\delta_n(d, F)$ and its asymptotic distribution $\delta_\infty(d, F)$ for (a) $d = \bar{\omega}$ and (b) $d = \bar{\zeta}_2$. The distribution F is exponential(1).



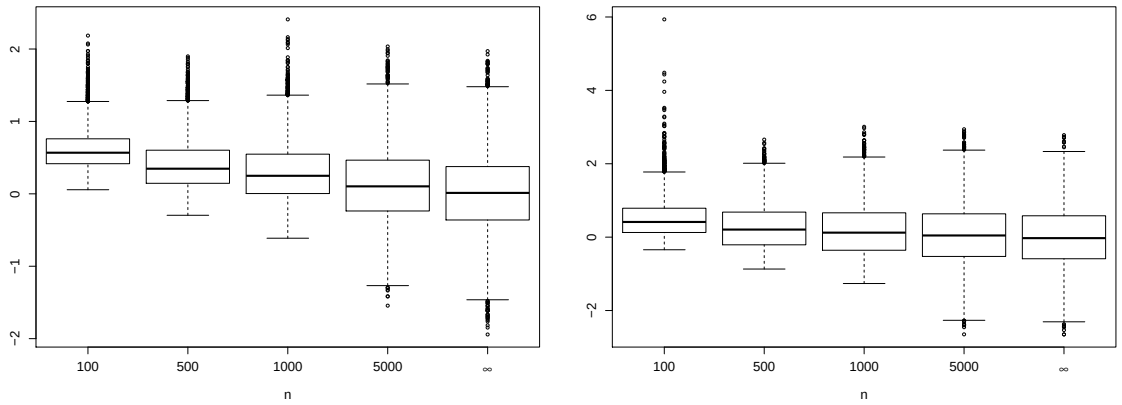
(a) (b)

Figure 5: Boxplots of $\delta_n(d, F)$ and its asymptotic distribution $\delta_\infty(d, F)$ for (a) $d = \bar{\omega}$ and (b) $d = \bar{\zeta}_2$. The distribution F is Weibull with shape parameter $a = 1.1$.



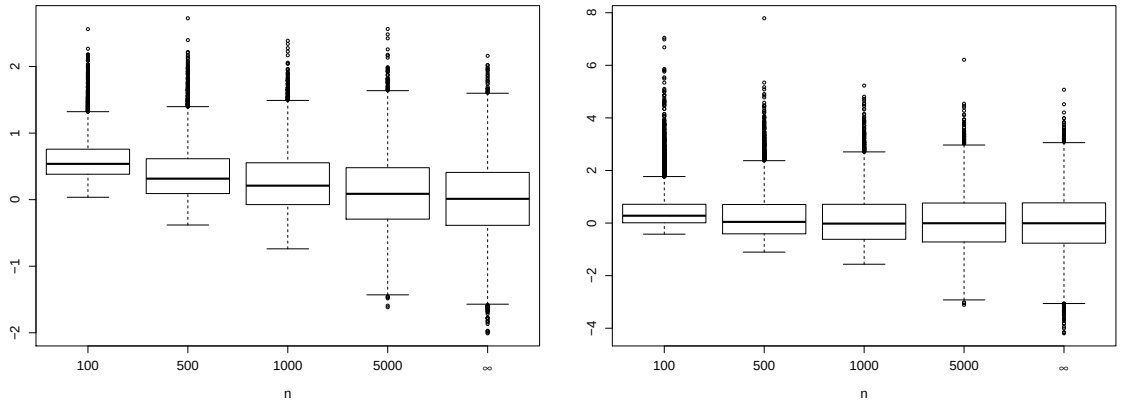
(a) (b)

Figure 6: Boxplots of $\delta_n(d, F)$ and its asymptotic distribution $\delta_\infty(d, F)$ for (a) $d = \bar{\omega}$ and (b) $d = \bar{\zeta}_2$. The distribution F is Weibull with shape parameter $a = 0.9$.



(a) (b)

Figure 7: Boxplots of $\delta_n(d, F)$ and its asymptotic distribution $\delta_\infty(d, F)$ for (a) $d = \bar{\omega}$ and (b) $d = \bar{\zeta}_2$. The distribution F is gamma with shape parameter $a = 1.1$.



(a) (b)

Figure 8: Boxplots of $\delta_n(d, F)$ and its asymptotic distribution $\delta_\infty(d, F)$ for (a) $d = \bar{\omega}$ and (b) $d = \bar{\zeta}_2$. The distribution F is gamma with shape parameter $a = 0.9$.

empirical normalized Wasserstein and Zolotarev distances, $\bar{\omega}(F_n, G_{\hat{\mu}})$ and $\bar{\zeta}_2(F_n, G_{\hat{\mu}})$, to the exponential class for the sample of PIT of each X-ray source in the COUP dataset, described in Sections 1 and 2. For the sake of comparison, we also include the usual Kolmogorov metric $\kappa(F_n, G_{\hat{\mu}})$ in the analysis as it is commonly used by astrophysicists in the classification of X-ray sources. The three types of sources of interest (namely, lightly-obscured PMS stars, heavily-obscured PMS stars and extragalactic) are compared via the values of these metrics.

For each of the X-ray sources in the three groups, we have computed the series of times between consecutive photon detections, taking into account the five observation gaps due to the passages through the high-radiation belts (see Section 2). We have also kept in mind that the complicated Chandra ACIS X-ray background, which is the combination of both celestial and instrumental backgrounds, tends to have peaks at energies below 0.5 keV and above 8 keV. To remove a significant portion of the X-ray background, we have followed [12], who apply an energy filter to the data by cleaning X-ray events with energies out of the total (0.5-8) keV band.

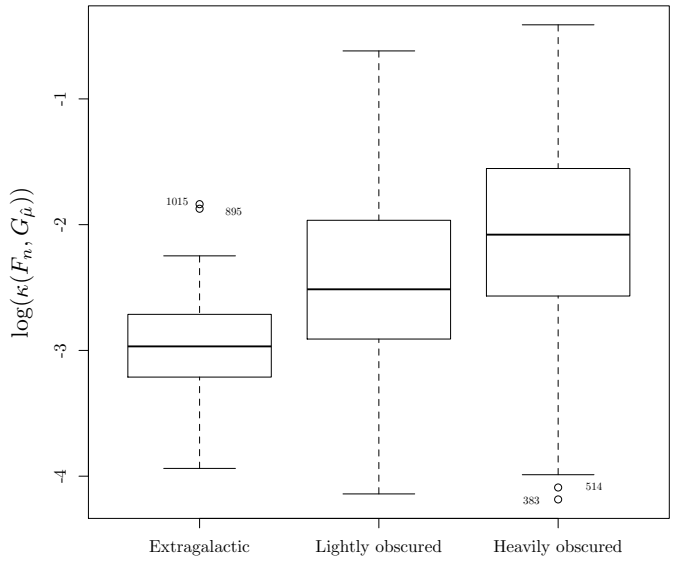
In order to obtain reasonably good estimates of the distances to the exponential distribution, we have kept only the COUP series with at least 100 PIT. We have finally analyzed 1090 samples of PIT, of which 73, 644 and 373 had been previously classified as extragalactic sources, lightly-obscured and heavily-obscured PMS stars, respectively. For each of these 1090 COUP sources, we have computed the distances $\kappa(F_n, G_{\hat{\mu}})$, $\bar{\omega}(F_n, G_{\hat{\mu}})$ and $\bar{\zeta}_2(F_n, G_{\hat{\mu}})$ of the empirical distribution of PIT to the exponential distribution with the same sample mean.

The interesting issue is whether the distance of PIT to the exponential distribution depends on the source class. The results are summarized in Figures 9 and 10. In Figure 9 we have displayed the boxplots of $\log(d(F_n, G_{\hat{\mu}}))$, for $d = \kappa$, $\bar{\omega}$, $\bar{\zeta}_2$, separated according to the three types of COUP sources. Outliers have been identified by the COUP source number (see [12]). As expected, for the three distances we see that the distribution of PIT due to extragalactic radiation is the nearest to the exponential distribution. Moreover, Figure 10 displays the empirical distribution functions of $\sqrt{n}\, d(F_n, G_{\hat{\mu}})$ ($d = \bar{\omega}$, $\bar{\zeta}_2$), for the different COUP groups and the distribution function of $\delta_\infty(d, G_1)$ when the distribution function is exponential with mean 1. The empirical distribution function corresponding to the extragalactic group is again the nearest to the distribution function of $\delta_\infty(d, G_1)$, the asymptotic distribution in Corollary 1. Additionally, Figure 9 shows that the normalized Zolotarev distance separates better the three classes than the normalized Wasserstein and the Kolmogorov-Smirnov metrics. Observe also that the PIT corresponding to lightly-obscured PMS stars are noticeably nearer to the exponential than those of the heavily-obscured group. A plausible explanation for this difference is the following: X-ray emission in heavily-obscured PMS stars is more affected by the intervening interstellar absorption, mainly the absorption from the gas in the molecular cloud and/or local absorption in an envelope or a disk around a star. Other factor might also be that the heavily absorbed sample is generally younger and more "diskier", in the sense that it has a higher fraction of stars that are still surrounded by their circumstellar disks; some of them highly accreting. Thus, the age and the presence/absence of disks could play an additional role, by affecting the X-ray production mechanisms of the strong X-ray flares in PMS stars

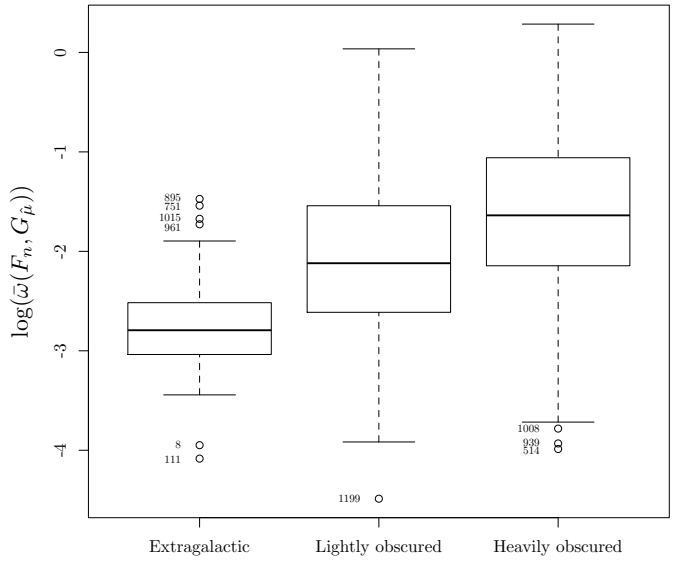
and in turn affecting their X-ray photon arrival patterns (see [13] and [14] for differences in the strong X-ray flares between the diskless and disky-accreting stellar populations).

The normalized metrics $\bar{\omega}$ and $\bar{\zeta}_2$ are also capable of detecting outlying COUP sources (see Figure 9), which are interesting to interpret. The X-ray light-curves and the photon arrival diagrams in the COUP atlas (see [12, Figure Set 12]), a numerical and graphical summary of all the COUP sources, show one thing in common among different types of outliers: the outlying sources nearest to the exponential distribution do not exhibit long and powerful X-ray flares, while all outliers farthest from 0 do. In other words, some outliers in Figure 9 could be in fact misclassified X-ray sources. For instance, the COUP sources 751, 895, 961, and 1015, classified as extragalactic sources by [11] and having the highest normalized Wasserstein and Zolotarev distances to the exponential distribution among those sources of the extragalactic sample (Figure 9), are likely to be young stellar objects in the Orion Nebula region.

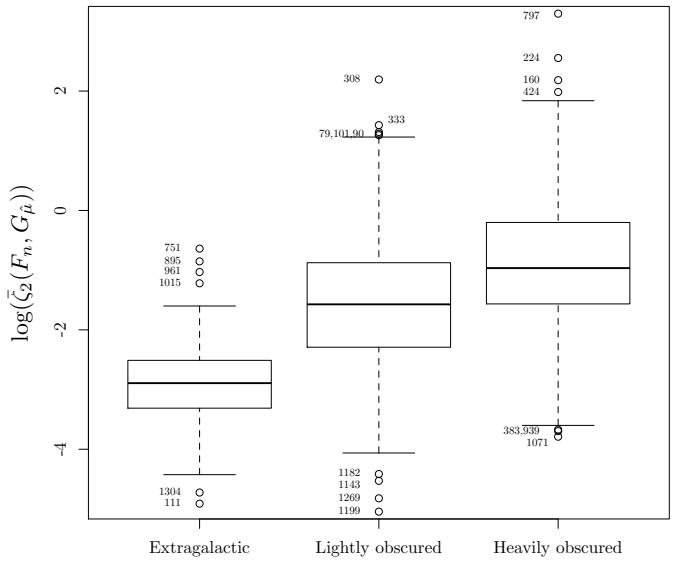
To analyze the nature of the above-mentioned four outliers with more precision, we decided to incorporate another informative variable with the potential of discriminating among the three COUP classes specified above. Taking into account that, for each of the photons belonging to a COUP source, we observe both the photon arrival time and its associated energy, it is natural to choose some feature summarizing the energies in the COUP series. The COUP atlas provides extra information (in the form of tabulated quantities for each COUP source). Specifically, we have considered the quantity MedEn, the median energy (in keV) of the source, due to its strong correlation with the absorbing column density characterizing interstellar absorption (see [9]). To further clarify why MedEn is an adequate choice, in Figure 11 we have plotted the empirical distribution functions of the photon energies rescaled to [0,1] for all the COUP sources. We clearly see that MedEn discriminates well between lightly obscured stars and extragalactic sources whereas distances to the exponential class separate better extragalactic sources from heavily obscured ones. Consequently, together these two quantities might separate well the three considered COUP classes, though, on their own, each of these variables fails to achieve a low misclassification rate. To check this, we have carried out three classification procedures on these data: quadratic discriminant analysis, the k-nearest neighbours (k-NN) rule and model-based discriminant analysis, respectively implemented in the R packages MASS ([28]), class ([28]) and mclust ([24]). We can choose k via cross-validation (CV). Using the package caret ([19]), we have run 10-, 5-fold and leave-one-out (LOO) CV to get insight into the value of $5 \leq k \leq 63$ yielding the largest accuracy. In the case of Zolotarev distance $\bar{\zeta}_2$ the optimal value of k is 5 for the three CV procedures. For the Wasserstein metric $\bar{\omega}$, the optimal k's are 35, 33 and 39 for 10-, 5-fold and LOO CV, respectively. For the Kolmogorov metric κ, the optimal k's are 19, 17 and 21 (or 23 and 25) for 10-, 5-fold and LOO CV, respectively. Since, for each metric, the optimal values of k are similar, we have chosen $k = 5$ for k-NN with the $\bar{\zeta}_2$ metric, $k = 35$ with $\bar{\omega}$ and $k = 19$ in the case of κ. The percentage of correct classifications with the three discriminant methods and the three distances appear in Table 1: they are all nearly the same and remarkably high (around 90%). The normalized Zolotarev distance of the PIT to the exponential distribution achieves the best correct classification rate, no matter which procedure is used. Although the improvement over the usual Kolmogorov metric may not seem significant, this is probably due to the low proportion of extragalactic sources in the sample. To



(a)



(b)



(c)

Figure 9: Analysis of COUP data. Boxplots of $\log(d(F_n, G_{\hat{\mu}}))$ for (a) $d = \kappa$, (b) $d = \bar{\omega}$ and (c) $d = \bar{\zeta}_2$.

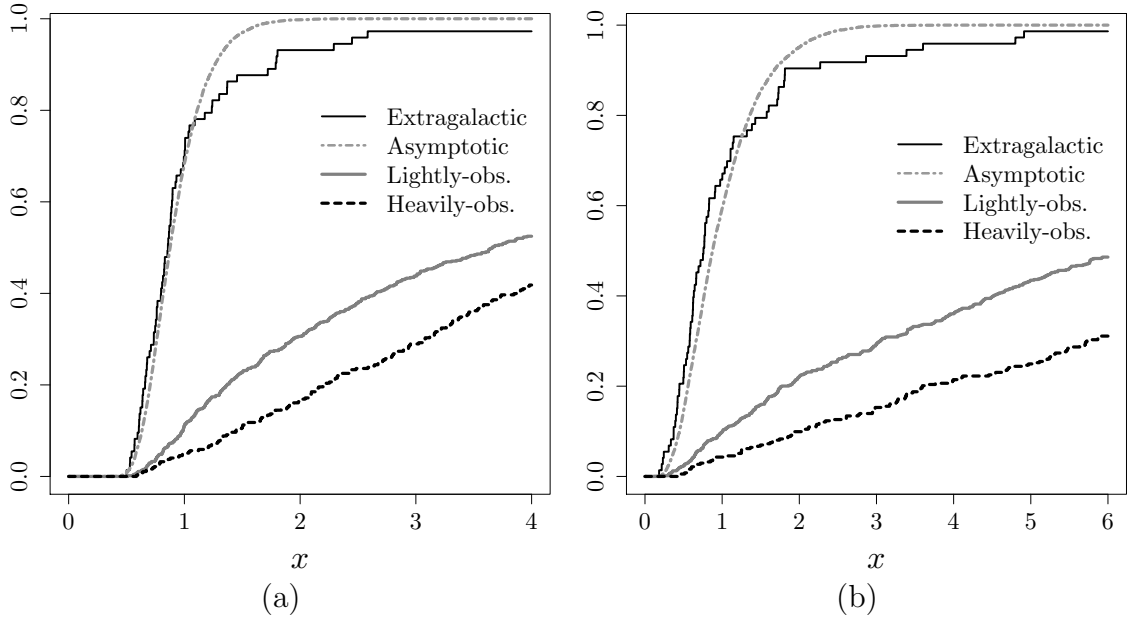


Figure 10: Analysis of COUP data. Distribution function of $\delta_\infty(d, G_1)$ (as given in Corollary 1) and empirical distribution functions of $\sqrt{n}\, d(F_n, G_{\hat{\mu}})$ for (a) $d = \bar{\omega}$ and (b) $d = \bar{\zeta}_2$.

gain a better insight of the advantages of the Zolotarev metric, in Table 2 we display the confusion matrices of these classification procedures. If we look at the extragalactic class (NM), using the Kolmogorov distance provides positive predictive values (PPVs) of 67%, 65% and 68%, for the quadratic, the k-NN and the mixture classification rules respectively. With the Zolotarev metric the corresponding PPVs are 78%, 76% and 78%, respectively. This significant increase in the classification accuracy was one of the motivations of this work.

There could be concerns regarding the potential influence of the outlying sources and their masking effect on the procedures employed in this section. For example, classical discriminant methods (like the quadratic or model-based rules) are sensitive towards outliers and may result in inaccurate parameter estimation or ill-posed problems. These methods can be replaced by robustified classifiers that rely, e.g., on depth functions ([17]) or on robust estimates of the mixture parameters ([10]). However, robust versions of the normalized distances of the PIT distributions to the exponential class are not adequate for this problem. The reason is that outlying interarrival times are precisely the ones leading to detection of flares, inconsistent with the usual behaviour of extragalactic radiation.

Of the three distances, κ, $\bar{\omega}$ and $\bar{\zeta}_2$, for the next diagram we have kept only the latter, as it has the highest discriminant ability. Figure 12 displays the scatterplot of the logarithm of MedEn in terms of the logarithm of the normalized Zolotarev distance to the exponential distribution. At a glance, we see the three COUP classes separated in clusters (extragalactic: high MedEn/low distance; heavily obscured: high energy/high distance; lightly obscured: low energy/medium distance).

Table 1: Analysis of COUP data: Percentage of correct classifications based on the logarithm of the median photon energy joint with the logarithm of the Kolmogorov distance (first row), the normalized Wasserstein distance (second row) and the normalized Zolotarev distance to the exponential distribution (last row).

Distance	Quadratic	k-NN	Model-based
Kolmogorov	88.99	87.89	89.08
Wasserstein	89.54	87.98	88.99
Zolotarev	90.18	90.09	90.64

Table 2: Analysis of COUP data: confusion matrices for classifications based on the logarithm of the median photon energy and the logarithm of the Kolmogorov distance (first row), the normalized Wasserstein distance (second row) and the normalized Zolotarev distance (last row) to the exponential distribution. Numbers in bold font correspond to the highest correct classification in each class. NM = non members, HO = Heavily obscured, LO = Lightly obscured stars.

Classification rule

Metric	**Quadratic**			**k-NN**			**Mixture**		
		Actual			Actual			Actual	
		NM	HO	LO	NM	HO	LO	NM	HO	LO
Kolmogorov Predicted	NM	48	23	1	50	26	1	48	22	1
	HO	25	299	20	23	292	27	25	308	28
	LO	0	51	623	0	55	616	0	43	615
		Actual			Actual			Actual	
		NM	HO	LO	NM	HO	LO	NM	HO	LO
Wasserstein Predicted	NM	55	21	2	54	26	2	54	22	2
	HO	18	297	18	18	283	20	19	295	21
	LO	0	55	**624**	1	64	**622**	0	56	**621**
		Actual			Actual			Actual	
		NM	HO	LO	NM	HO	LO	NM	HO	LO
Zolotarev Predicted	NM	**59**	15	2	**57**	15	3	**60**	15	2
	HO	14	**302**	20	16	**317**	33	13	**307**	21
	LO	0	56	622	0	41	608	0	51	**621**

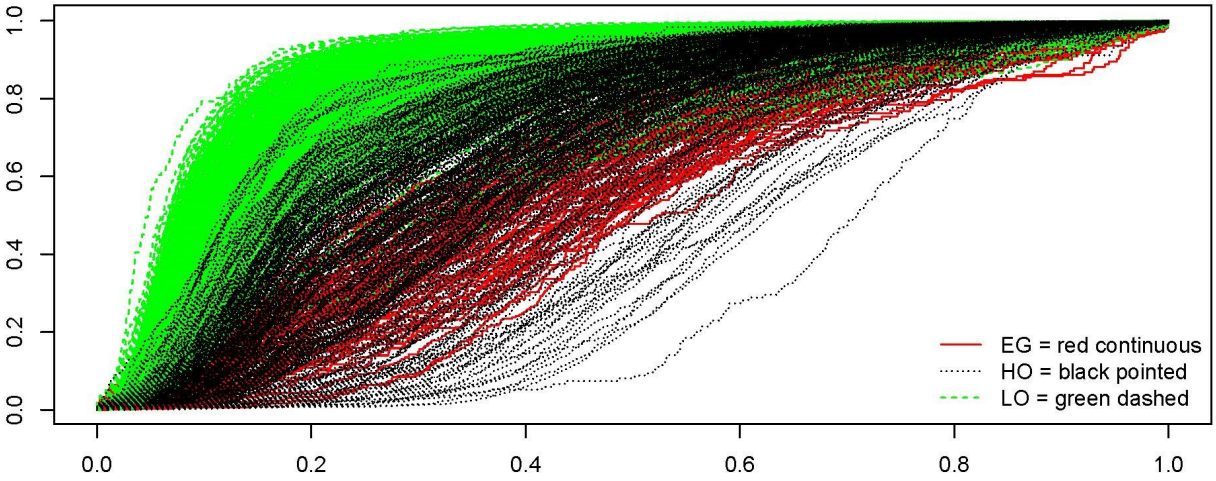


Figure 11: Analysis of COUP data: Empirical distribution functions of the photon energies rescaled to the interval [0,1].

Table 3: Analysis of COUP data: Posterior probabilities of membership for extragalactic outliers.

	Posterior probabilities (in percentage)		
Source	Extragalactic	Heavily obs.	Lightly obs.
751	0.0152	92.8997	7.08514
895	1.5713	98.4230	0.00576
961	0.6601	98.9164	0.42343
1015	8.4486	91.5511	0.00003

In Figure 12, a simple visual inspection reveals that COUP sources 751, 895 and 961, classified as extragalactic, have a higher probability of being heavily obscured stars. Indeed, Table 3 displays the posterior probabilities of membership, derived from the quadratic classification rule, for the misclassified extragalactic cases when these probabilities exceed 0.9. We see that all the sources in Table 3 are actually the largest outliers "detected" by the normalized Zolotarev distance $\bar{\zeta}_2$ in the extragalactic class (Figure 9(c)). This emphasizes the information conveyed by $\bar{\zeta}_2$ on the source class.

6 Conclusions

We have introduced normalized versions of two integral probability metrics, the Wasserstein and Zolotarev distances, to quantify the discrepancy between photon interarrival times of X-ray cosmic sources and the exponential distribution. The aim is to measure how different the photon emitting source is from extragalactic radiation. The plug-in estimators of these metrics show a good asymptotic behaviour. The analysis of more than one thousand X-ray sources from the Chandra Orion Ultradeep Project with the proposed metrics reveals that the information conveyed by photon interarrival times on the nature of each X-ray source

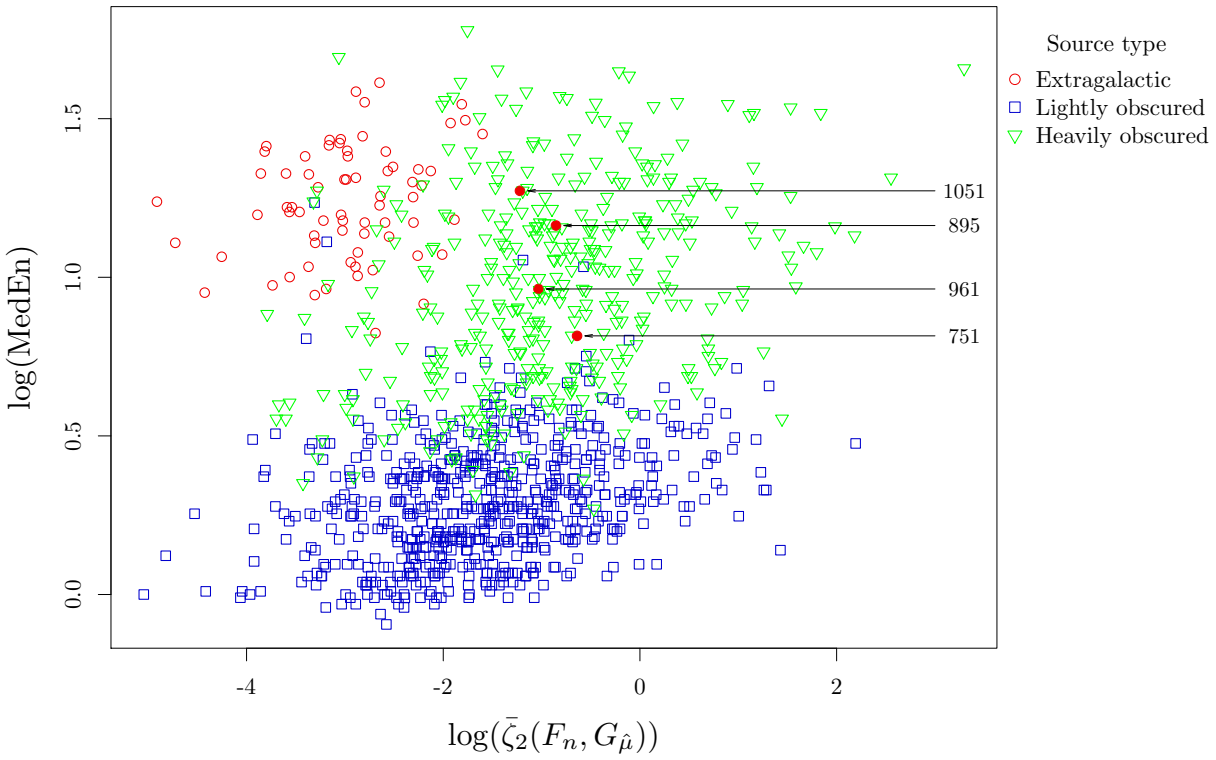


Figure 12: Analysis of COUP data: Scatterplot of the logarithm of the median energy and the logarithm of the normalized Zolotarev distance to the exponential distribution. The four outlying sources of the extragalactic class, identified by the Zolotarev distance, are highlighted with filled circles.

is very well summarized by the normalized Zolotarev distance. We have further shown that this metric, together only with the median energy of the X-ray source, yields a high percentage of correct classifications of the sources into the classes previously provided by astrophysicists. We remark that here we have only used two discriminating features while the usual expert procedures in astronomy rely on many more. As a striking conclusion, we have detected four sources, originally classified as extragalactic, which, as a matter of fact, are very likely young stars in Orion Molecular Cloud 1.

Appendix

In this technical appendix, we collect the proofs of the results in Section 3.2. The main ingredients are the following: first, we show that the sequences of stochastic processes defined in (7) are equivalent in L^1 to continuous functionals of the empirical process; then we apply the central limit theorem (CLT) in suitable Banach spaces to find their weak limits; finally, we use the continuity of the functional given in (6) to derive the asymptotic distribution of $\delta_n(d, F)$ in (4).

To begin, we recall that if the stochastic processes $\mathbb{P}_n$ and $\mathbb{P}$ take values in L^1 a.s., it is said that $\mathbb{P}_n$ *converges in distribution to* $\mathbb{P}$ *in* L^1 if $\lim_{n\to\infty} \mathrm{E} f(\mathbb{P}_n) = \mathrm{E} f(\mathbb{P})$, for all continuous and bounded functions $f : L^1 \longrightarrow \mathbb{R}$. Note that if $\mathbb{P}$ and $\mathbb{P}_n$ are jointly measurable and have almost all their trajectories in L^1, they can be identified with Borel-measurable random elements in L^1. Therefore, the previous expectations are well-defined. In the following, we denote this weak convergence of probability measures in L^1 by $\mathbb{P}_n \xrightarrow{L^1}_{\mathrm{w}} \mathbb{P}$. An analogous definition can be given for the weak convergence in the weighted L^1 space defined by

$$W^1 := \left\{ f \in L^1 : \|f\|_{W^1} := \int_0^\infty (1+t)|f(t)|\,\mathrm{d}t < \infty \right\}.$$

Let $\mathbb{P}_n$ and $\widetilde{\mathbb{P}}_n$ be two stochastic processes with trajectories in L^1 a.s. We say that $\mathbb{P}_n$ and $\widetilde{\mathbb{P}}_n$ are *equivalent in* L^1, denoted by $\mathbb{P}_n \overset{L^1}{\sim} \widetilde{\mathbb{P}}_n$, if $\|\mathbb{P}_n - \widetilde{\mathbb{P}}_n\|_1 \longrightarrow_{\mathrm{P}} 0$, where "$\longrightarrow_{\mathrm{P}}$" stands for convergence in probability. Roughly speaking, if $\mathbb{P}_n \overset{L^1}{\sim} \widetilde{\mathbb{P}}_n$, the two processes have the same asymptotic behavior in L^1 because if $\mathbb{P}_n \xrightarrow{L^1}_{\mathrm{w}} \mathbb{P}$ and $\mathbb{P}_n \overset{L^1}{\sim} \widetilde{\mathbb{P}}_n$, then $\widetilde{\mathbb{P}}_n \xrightarrow{L^1}_{\mathrm{w}} \mathbb{P}$ (see for instance [**?**, Theorem 18.10]).

In the sequel, $\mathbb{E}_n$ stands for the *empirical process* associated to X, that is, $\mathbb{E}_n(t) := \sqrt{n}(F_n(t) - F(t))$, $t \geq 0$, $n \geq 1$. The asymptotic behavior of $\mathbb{E}_n$ in L^1 and W^1 is collected in the following lemma. Part (a) is a known result (see [**?**, Theorem 2.1]), while part (b) can be found in [3, Lemma 2].

Lemma 1. *We have that*

(a) $\mathbb{E}_n \xrightarrow{L^1}_{\mathrm{w}} \mathbb{B}_F$ *if and only if* $X \in \mathcal{L}^{2,1}$.

(b) $\mathbb{E}_n \xrightarrow{W^1}_{w} \mathbb{B}_F$ *if and only if* $X \in \mathcal{L}^{4,2}$.

Our first task is to find processes expressed as continuous functionals of $\mathbb{E}_n$ and equivalent to $\mathbb{X}_{d,n}$ in L^1. To start, for $t \geq 0$, we decompose $\mathbb{X}_{d,n}$ given in (7) as:

$$\mathbb{X}_{d,n} = \mathbb{A}_{d,n} + \mathbb{B}_{d,n} + \mathbb{C}_{d,n}, \tag{15}$$

where

$$\begin{aligned}
&\mathbb{A}_{\omega,n} := \mathbb{E}_n, && \mathbb{B}_{\omega,n} := \sqrt{n}(G_\mu - G_{\hat{\mu}}), && \mathbb{C}_{\omega,n} := 0,\\
&\mathbb{A}_{\bar{\omega},n} := \mathbb{E}_n/\hat{\mu}, && \mathbb{B}_{\bar{\omega},n} := \sqrt{n}(G_\mu - G_{\hat{\mu}})/\hat{\mu}, && \mathbb{C}_{\bar{\omega},n} := g_{\bar{\omega}}\sqrt{n}(\mu - \hat{\mu})/\hat{\mu},\\
&\mathbb{A}_{\zeta_2,n}(t) := \int_t^\infty \mathbb{E}_n, && \mathbb{B}_{\zeta_2,n}(t) := \sqrt{n}\int_t^\infty (G_\mu - G_{\hat{\mu}}), && \mathbb{C}_{\zeta_2,n} := 0,\\
&\mathbb{A}_{\bar{\zeta}_2,n}(t) := \frac{1}{\hat{\mu}^2}\int_t^\infty \mathbb{E}_n, && \mathbb{B}_{\bar{\zeta}_2,n}(t) := \frac{\sqrt{n}}{\hat{\mu}^2}\int_t^\infty (G_\mu - G_{\hat{\mu}}), && \mathbb{C}_{\bar{\zeta}_2,n} := g_{\bar{\zeta}_2}\frac{\sqrt{n}(\mu^2 - \hat{\mu}^2)}{\hat{\mu}^2},
\end{aligned}$$

with $g_{\bar{\omega}}$ and $g_{\bar{\zeta}_2}$ defined in (8) and (9), respectively.

Lemma 2 provides equivalent expressions for the processes defined above.

Lemma 2. *Let X be a positive random variable with mean $\mu > 0$. For $t \geq 0$, the following assertions hold:*

(a) If $X \in \mathcal{L}^{4/3}$, $\mathbb{B}_{\omega,n} \overset{L^1}{\sim} \widetilde{\mathbb{B}}_{\omega,n}$, where $\widetilde{\mathbb{B}}_{\omega,n}(t) := \sqrt{n}(\hat{\mu} - \mu)te^{-t/\mu}/\mu^2$.

(b) If $X \in \mathcal{L}^{2,1}$, $\mathbb{A}_{\bar{\omega},n} \overset{L^1}{\sim} \widetilde{\mathbb{A}}_{\bar{\omega},n} := \mathbb{E}_n/\mu$.

(c) If $X \in \mathcal{L}^{4/3}$, $\mathbb{B}_{\bar{\omega},n} \overset{L^1}{\sim} \widetilde{\mathbb{B}}_{\bar{\omega},n}$, where $\widetilde{\mathbb{B}}_{\bar{\omega},n}(t) := \sqrt{n}(\hat{\mu} - \mu)te^{-t/\mu}/\mu^3$.

(d) If $X \in \mathcal{L}^{4/3}$, $\mathbb{C}_{\bar{\omega},n} \overset{L^1}{\sim} \widetilde{\mathbb{C}}_{\bar{\omega},n} := \sqrt{n}(\mu - \hat{\mu})g_{\bar{\omega}}/\mu$.

(e) If $X \in \mathcal{L}^{4/3}$, $\mathbb{B}_{\zeta_2,n} \overset{L^1}{\sim} \widetilde{\mathbb{B}}_{\zeta_2,n}$, where $\widetilde{\mathbb{B}}_{\zeta_2,n}(t) := \sqrt{n}(\hat{\mu} - \mu)(1 + t/\mu)e^{-t/\mu}$.

(f) If $X \in \mathcal{L}^{4,2}$, $\mathbb{A}_{\bar{\zeta}_2,n} \overset{L^1}{\sim} \widetilde{\mathbb{A}}_{\bar{\zeta}_2,n}$, where $\widetilde{\mathbb{A}}_{\bar{\zeta}_2,n}(t) := \int_t^\infty \mathbb{E}_n/\mu^2$.

(g) If $X \in \mathcal{L}^{4/3}$, $\mathbb{B}_{\bar{\zeta}_2,n} \overset{L^1}{\sim} \widetilde{\mathbb{B}}_{\bar{\zeta}_2,n}$, where $\widetilde{\mathbb{B}}_{\bar{\zeta}_2,n}(t) := \sqrt{n}(\hat{\mu} - \mu)(1 + t/\mu)e^{-t/\mu}/\mu^2$.

(h) If $X \in \mathcal{L}^2$, $\mathbb{C}_{\bar{\zeta}_2,n} \overset{L^1}{\sim} \widetilde{\mathbb{C}}_{\bar{\zeta}_2,n} := \sqrt{n}(\mu - \hat{\mu})2\, g_{\bar{\zeta}_2}/\mu$.

Proof. To show part (a), we use the mean value theorem twice to obtain

$$\|\mathbb{B}_{\omega,n} - \widetilde{\mathbb{B}}_{\omega,n}\|_1 \leq \sqrt{n}(\hat{\mu} - \mu)^2 \int_0^\infty t|2 - t/\xi_t|e^{-t/\xi_t}/\xi_t^3 \,\mathrm{d}t, \tag{16}$$

where ξ_t is a point between μ and $\hat{\mu}$. The integral in (16) is bounded by

$$\int_0^\infty t|2 - t/\xi_t|e^{-t/\xi_t}/\xi_t^3 \,\mathrm{d}t \leq 2(\mu + \hat{\mu})\frac{\max\{\mu, \hat{\mu}\}^2}{\min\{\mu, \hat{\mu}\}^4} \to 4/\mu \quad \text{a.s.} \tag{17}$$

Therefore, from (16)-(17), and by the Kolmogorov, Marcinkiewicz and Zygmund strong law of large numbers (see, e.g., [18, Theorem 3.23]), we see that, whenever $X \in \mathcal{L}^{4/3}$, $\|\mathbb{B}_{\omega,n} - \widetilde{\mathbb{B}}_{\omega,n}\|_1 \to 0$ a.s.

To see (b), we note that $\|\mathbb{A}_{\bar{\omega},n} - \widetilde{\mathbb{A}}_{\bar{\omega},n}\|_1 = \|\mathbb{E}_n\|_1 |\mu - \hat{\mu}|/(\mu\hat{\mu})$. From Lemma 1 (a), we have that $\|\mathbb{E}_n\|_1 \longrightarrow_{\mathrm{d}} \|\mathbb{B}_F\|_1$ and the conclusion follows from the strong law of large numbers and Slutsky's theorem.

Part (c) follows from (a), as it is straightforward to check that $\mathbb{B}_{\bar{\omega},n} \overset{L^1}{\sim} \mathbb{B}_{\omega,n}/\mu$, whenever $X \in \mathcal{L}^{4/3}$.

Part (d) is direct, whereas part (e) can be found in [3, Lemma 1]. The proof of part (f) is similar to the one for (b) by using Lemma 1 (b).

To show part (g), we observe that, from part (e), we have that $\widetilde{\mathbb{B}}_{\bar{\zeta}_2,n} \overset{L^1}{\sim} \mathbb{B}_{\zeta_2,n}/\mu^2$. The conclusion follows by checking that $\|\mathbb{B}_{\bar{\zeta}_2,n} - \mathbb{B}_{\zeta_2,n}/\mu^2\|_1 = \sqrt{n}(\mu - \hat{\mu})^2(1/\mu + 1/\hat{\mu})^2$.

Finally, it can be seen that $\|\mathbb{C}_{\bar{\zeta}_2,n} - \widetilde{\mathbb{C}}_{\bar{\zeta}_2,n}\|_1 = \sqrt{n}(\mu - \hat{\mu})^2(1/\mu + 1/\hat{\mu})\, \bar{\zeta}_2(X, Y_\mu)$. As $\bar{\zeta}_2(X, Y_\mu) < \infty$ if and only if $X \in \mathcal{L}^2$, we conclude that (h) is fulfilled. □ □

The next corollary, which is a consequence of Lemma 2 and (15), shows that $\mathbb{X}_{d,n}$ are equivalent in L^1 to certain continuous functionals of the empirical process $\mathbb{E}_n$.

Corollary 3. *Let X be a positive random variable with mean $\mu > 0$.*

(i) If $X \in \mathcal{L}^{4/3}$, then $\mathbb{X}_{\omega,n} \overset{L^1}{\sim} \phi_\omega(\mathbb{E}_n)$, where $\phi_\omega : L^1 \to L^1$ is the linear operator

$$\phi_\omega(f,t) := f(t) - \frac{t}{\mu^2} e^{-t/\mu} \int_0^\infty f(x)\,\mathrm{d}x, \quad t \geq 0.$$

Moreover, $\|\phi_\omega(f)\|_1 \leq 2\,\|f\|_1$, and ϕ_ω is therefore continuous.

(ii) If $X \in \mathcal{L}^{2,1}$, then $\mathbb{X}_{\bar{\omega},n} \overset{L^1}{\sim} \phi_{\bar{\omega}}(\mathbb{E}_n)$, where $\phi_{\bar{\omega}} : L^1 \to L^1$ is the linear operator

$$\phi_{\bar{\omega}}(f,t) := \frac{1}{\mu}\left[f(t) + \left(g_{\bar{\omega}}(t) - \frac{t}{\mu^2} e^{-t/\mu}\right) \int_0^\infty f(x)\,\mathrm{d}x\right], \quad t \geq 0.$$

Moreover, $\|\phi_{\bar{\omega}}(f)\|_1 \leq \|f\|_1 (2 + \bar{\omega}(X, Y_\mu))/\mu$, and $\phi_{\bar{\omega}}$ is therefore continuous.

(iii) If $X \in \mathcal{L}^{4/3}$, then $\mathbb{X}_{\zeta_2,n} \overset{L^1}{\sim} \phi_{\zeta_2}(\mathbb{E}_n)$, where $\phi_{\zeta_2} : W^1 \to L^1$ is the linear operator

$$\phi_{\zeta_2}(f,t) := \int_t^\infty f(x)\,\mathrm{d}x - \left(1 + \frac{t}{\mu}\right) e^{-t/\mu} \int_0^\infty f(x)\,\mathrm{d}x, \quad t \geq 0.$$

Moreover, $\|\phi_{\zeta_2}(f)\|_1 \leq (1 + 2\mu)\,\|f\|_{W^1}$, and ϕ_{ζ_2} is therefore continuous.

(iv) If $X \in \mathcal{L}^{4,2}$, then $\mathbb{X}_{\bar{\zeta}_2,n} \overset{L^1}{\sim} \phi_{\bar{\zeta}_2}(\mathbb{E}_n)$, where $\phi_{\bar{\zeta}_2} : W^1 \to L^1$ is the linear operator

$$\phi_{\bar{\zeta}_2}(f,t) := \frac{1}{\mu^2}\left[\int_t^\infty f(x)\,\mathrm{d}x + \left(2\,\mu\, g_{\bar{\zeta}_2}(t) - \left(1 + \frac{t}{\mu}\right) e^{-t/\mu}\right) \int_0^\infty f(x)\,\mathrm{d}x\right],$$

for $t \geq 0$. Moreover, $\|\phi_{\bar{\zeta}_2}(f)\|_1 \leq \|f\|_{W^1}\,[1 + 2\mu(1 + \bar{\zeta}_2(X, Y_\mu))]/\mu^2$, and $\phi_{\bar{\zeta}_2}$ is therefore continuous.

We are now in condition to prove Theorems 1 and 2.

Proof of Theorem 1 Assume that (a) holds, i.e., $X \in \mathcal{L}^{2,1}$. For $d = \omega$ or $d = \bar{\omega}$, from Lemma 1 (a) and Corollary 3 (i) and (ii), we have that $\mathbb{X}_{d,n} \overset{L^1}{\sim} \phi_d(\mathbb{E}_n) \xrightarrow{L^1}_{\mathrm{w}} \phi_d(\mathbb{B}_F) = \mathbb{X}_{d,F}$, by the continuos mapping theorem. We conclude that (b) and (c) hold.

Conversely, let us assume that (b) is satisfied. By Corollary 3 (i), if $X \in \mathcal{L}^{4/3}$, we obtain that $\phi_\omega(\mathbb{E}_n) \xrightarrow{L^1}_{\mathrm{w}} \mathbb{X}_{\omega,F}$. Observe now that $\phi_\omega(\mathbb{E}_n)$ can be rewritten as the normalized sum $\phi_\omega(\mathbb{E}_n) = n^{-1/2} \sum_{i=1}^n \mathbb{Y}_{\omega,i}$ where $\mathbb{Y}_{\omega,1}, \ldots, \mathbb{Y}_{\omega,n}$ are n independent copies of the zero mean process

$$\mathbb{Y}_\omega(t) := \mathrm{P}(X > t) - I_{\{X>t\}} + (X - \mu)te^{-t/\mu}/\mu^2, \quad t \geq 0. \tag{18}$$

This means that the process $\mathbb{Y}_\omega$ satisfies the CLT in L^1 (and implies that $\mathbb{X}_{\omega,F}$ is a centered Gaussian process), which is equivalent (see [1, p. 205]) to

$$\int_0^\infty \sqrt{\mathrm{E}\mathbb{Y}_\omega(t)^2}\, \mathrm{d}t < \infty. \tag{19}$$

In particular, this implies that $X \in \mathcal{L}^2$ and denoting $\mathbb{Z}(t) := \mathrm{P}(X > t) - I_{\{X>t\}}$ $(t \geq 0)$, from (18), (19), and by Minkowski inequality, we have that

$$\int_0^\infty \sqrt{\mathrm{E}\mathbb{Z}(t)^2}\, \mathrm{d}t \leq \sigma + \int_0^\infty \sqrt{\mathrm{E}\mathbb{Y}_\omega(t)^2}\, \mathrm{d}t < \infty,$$

where σ is the standard deviation of X. Last inequality amounts to (a) $X \in \mathcal{L}^{2,1}$.

For the proof that (c) implies (a), it is enough to note that $\mathbb{X}_{\bar{\omega},n} \xrightarrow{L^1}_{\mathrm{w}} \mathbb{X}_{\bar{\omega},F}$ is equivalent to $\hat{\mu}\, \mathbb{X}_{\bar{\omega},n} \xrightarrow{L^1}_{\mathrm{w}} \mu\, \mathbb{X}_{\bar{\omega},F}$. The rest of the proof runs as with $d = \omega$. □

Proof of Theorem 2 To show that part (a) implies (b) and (c) it is enough to follow the same steps as in the proof of the same implications in Theorem 1. We omit the details.

To finish, we will show that part (c) implies (a) (the remaining implication "(b) ⇒ (a)" is simpler and similar). Let us assume that (c) is satisfied. In this situation, it is clear that $X \in \mathcal{L}^2$, as this integrability condition amounts to saying that the process $\mathbb{X}_{\bar{\zeta}_2,n}$ has its paths in L^1 a.s. We have that $\hat{\mu}^2 \mathbb{X}_{\bar{\zeta}_2,n} \xrightarrow{L^1}_{\mathrm{w}} \mu^2 \mathbb{X}_{\bar{\zeta}_2,F}$. Further, by Lemma 2, we conclude that $\hat{\mu}^2 \mathbb{X}_{\bar{\zeta}_2,n} \overset{L^1}{\sim} \overline{\mathbb{Z}}_n$, where

$$\overline{\mathbb{Z}}_n(t) := \int_t^\infty \mathbb{E}_n + h(t)\sqrt{n}(\mu - \hat{\mu}), \quad t \geq 0,$$

with

$$h(t) := 2\mu\, g_{\bar{\zeta}_2}(t) - (1 + t/\mu)e^{-t/\mu}, \quad t \geq 0. \tag{20}$$

The process $\overline{\mathbb{Z}}_n$ can be rewritten as the normalized sum $\overline{\mathbb{Z}}_n = n^{-1/2} \sum_{i=1}^n \mathbb{Z}_{\bar{\zeta}_2,i}$, where $\mathbb{Z}_{\bar{\zeta}_2,1}, \ldots, \mathbb{Z}_{\bar{\zeta}_2,n}$ are n independent copies of the zero mean process

$$\mathbb{Z}_{\bar{\zeta}_2}(t) := \mathrm{E}(X - t)_+ - (X - t)_+ + h(t)\,(\mu - X).$$

Therefore, the process $\mathbb{Z}_{\bar{\zeta}_2}$ satisfies the CLT in L^1 (and we see that $\mathbb{X}_{\bar{\zeta}_2,F}$ is a centered Gaussian process). Using again [1, p. 205], we obtain that

$$\int_0^\infty \sqrt{\mathrm{E}\mathbb{Z}_{\bar{\zeta}_2}(t)^2}\,\mathrm{d}t < \infty. \tag{21}$$

Finally, by Minkowski inequality and Fubini theorem, we have that

$$\int_0^\infty \sqrt{\mathrm{E}(X-t)_+^2}\,\mathrm{d}t \leq \mathrm{E}X^2/2 + \sigma\|h\|_1 + \int_0^\infty \sqrt{\mathrm{E}\mathbb{Z}_{\bar{\zeta}_2}(t)^2}\,\mathrm{d}t.$$

From (20), we can check that $\|h\|_1 \leq 2\mu(1+\bar{\zeta}_2(X,Y_\mu)) < \infty$, and from (21), we conclude that $\int_0^\infty \sqrt{\mathrm{E}(X-t)_+^2}\,\mathrm{d}t < \infty$. This implies that $X \in \mathcal{L}^{4,2}$ because $t^2\mathrm{P}(X > 2t) \leq \mathrm{E}(X-t)_+^2$. □

Proof of Theorem 3 We have that $\delta_n(d,F) = \rho_n(\mathbb{X}_{d,n}, g_d)$, with ρ_n defined in (6). The continuity of ρ_n (in L^1) was analyzed in [6, Lemma 4], where it was shown that if $f_n \to f$ in L^1 and $g \in L^1$, then, for $I(g) := \{t \geq 0 : g(t) = 0\}$,

$$\lim_{n\to\infty} \rho_n(f_n, g) = \rho(f,g) := \int_{I(g)} |f| + \int_{I(g)^c} f\,\mathrm{sgn}(g). \tag{22}$$

Therefore, from Theorems 1 and 2, (22), and the extended continuous mapping theorem (see [**?**, Theorem 1.11.1]), we conclude that $\delta_n(d,F) = \rho_n(\mathbb{X}_{d,n}, g_d) \longrightarrow_{\mathrm{d}} \rho(\mathbb{X}_{d,F}, g_d) = \delta_\infty(d,F)$, as $n \to \infty$. □

Proof of Corollary 2 From Theorem 3, we have that

$$\delta_\infty(d,F) = \int_0^\infty \mathbb{X}_{d,F}(t)\,\mathrm{sgn}(g_d(t))\,\mathrm{d}t.$$

As $\mathbb{X}_{d,F}$ is a centered Gaussian process and g_d is nonrandom, we conclude that $\delta_\infty(d,F)$ is normally distributed. □

Acknowledgements

The authors are grateful to three reviewers and the associate editor for their insightful comments which have improved the presentation of the paper.

References

[1] Araujo, A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables*. Wiley.

[2] Ascher, S. (1990). A survey of tests for exponentiality. *Communications in Statistics-Theory and Methods*, 19(5), 1811–1825.

[3] Baíllo, A., Cárcamo, J. and Nieto, E. (2015). A test for convex dominance with respect to the exponential class based on an L^1 distance. *IEEE Transactions on Reliability*, 64, 71–82.

[4] del Barrio, E., Giné, E. and Matrán, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *The Annals of Probability*, 27, 1009–1071.

[5] Broos, P.S., Getman, K.V., Povich, M.S., Townsley, L.K., Feigelson, E.D. and Garmire, G.P. (2011). A naive Bayes source classifier for X-ray sources. *The Astrophysical Journal Supplement Series* 194, 4.

[6] Cárcamo, J. (2017). Integrated empirical processes in L^p with applications to estimate probability metrics. *Bernoulli*, 23(4B), 3412–3436.

[7] Efron, B., and Tibshirani, R.J. (1993). *An introduction to the bootstrap*. CRC press.

[8] Feigelson, E.D. and Babu, G.J. (2012). *Modern Statistical Methods for Astronomy. With R Applications*. Cambridge University Press.

[9] Feigelson, E.D., Getman, K., Townsley, L., Garmire, G., Preibisch, T., Grosso, N., Montmerle, T., Muench, A. and McCaughrean, M. (2005). Global X-ray properties of the Orion Nebula region. *The Astrophysical Journal Supplement Series*, 160, 379–389.

[10] García-Escudero, L.A., Gordaliza, A. and Mayo-Iscar, A. (2014). A constrained robust proposal for mixture modeling avoiding spurious solutions. *Advances in Data Analysis and Classification*, 8, 27–43.

[11] Getman, K.V., Feigelson, E.D., Grosso, N., McCaughrean, M.J., Micela, G., Broos, P. Garmire, G. and Townsley, L. (2005a). Membership of the Orion Nebula population from the Chandra Orion Ultradeep Project. *The Astrophysical Journal Supplement Series*, 160, 353–378.

[12] Getman, K.V., Flaccomio, E., Broos, P.S., Grosso, N., Tsujimoto, M., Townsley, L., Garmire, G.P., Kastner, J., Li, J., Harnden, F.R., Wolk, S., Murray, S.S., Lada, C.J., Muench, A.A., McCaughrean, M.J., Meeus, G., Damiani, F., Micela, G., Sciortino, S., Bally, J., Hillenbrand, L.A., Herbst, W., Preibisch, T. and Feigelson, E. (2005b). Chandra Orion Ultradeep Project: Observations and source lists. *The Astrophysical Journal Supplement Series*, 160, 319–352.

[13] Getman, K.V., Feigelson, E.D., Broos, P.S., Micela, G. and Garmire, G.P. (2008a). X-ray flares in Orion young stars. I. Flare characteristics. *The Astrophysical Journal*, 688, 418–436.

[14] Getman, K.V., Feigelson, E.D., Micela, G., Jardine, M.M., Gregory, S.G. and Garmire, G.P. (2008b). X-ray flares in Orion young stars. II. Flares, magnetospheres, and protoplanetary disks. *The Astrophysical Journal*, 688, 437–455.

[15] Grafakos, L. (2014). *Classical Fourier Analysis*. Springer.

[16] Henze, N., and Meintanis, S.G. (2005). Recent and classical tests for exponentiality: a partial review with comparisons. *Metrika*, 61(1), 29–45.

[17] Hubert, M., Rousseeuw, P. and Segaert, P. (2017). Multivariate and functional classification using depth and distance. *Advances in Data Analysis and Classification*, 11, 445–466.

[18] Kallemberg, O. (1997). *Foundations of Modern Probability*. Springer.

[19] Kuhn, M. Contributions from J. Wing, S. Weston, A. Williams, C. Keefer, A. Engelhardt, T. Cooper, Z. Mayer, B. Kenkel, the R Core Team, M. Benesty, R. Lescarbeau, A. Ziem, L. Scrucca, Y. Tang, C. Candan and T. Hunt. (2016). caret: Classification and Regression Training. R package version 6.0-72. `https://CRAN.R-project.org/package=caret`

[20] Ledoux, M. and Talagrand, M. (2002). *Probability in Banach Spaces*. Springer.

[21] Rachev, S.T., Klebanov, L., Stoyanov, S.V. and Fabozzi, F. (2013). *The Methods of Distances in the Theory of Probability and Statistics*. Springer.

[22] Rachev, S.T., Stoyanov, S.V. and Fabozzi, F. (2011). *A Probability Metrics Approach to Financial Risk Measures*. Wiley-Blackwell.

[23] Schulz, N.S. (2012). *The Formation and Early Evolution of Stars: From Dust to Stars and Planets*. Second edition. Springer.

[24] Scrucca L., Fop M., Murphy T.B.and Raftery A.E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *R Journal*, 8, 289–317.

[25] Shaked, M. and Shanthikumar, J.G. (2006). *Stochastic Orders*. Springer.

[26] van der Vaart, A. (1998). *Asymptotic Statistics*. Cambridge University Press.

[27] van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer.

[28] Venables, W.N., and Ripley, B.D. (2002). *Modern Applied Statistics with S*. Springer, New York.

[29] Wolk, S.J., Harnden, F.R., Jr, Flaccomio, E., Micela, G., Favata, F., Shang, H. and Feigelson, E.D. (2005). Stellar activity on the young suns of Orion: COUP observations of K5-7 pre-main-sequence stars. *The Astrophysical Journal Supplement Series* 160, 423–449.