Author Name

Data Mining and Machine Learning for Astronomical Applications

Contents

I This is a Part **1**

1 Time-Frequency Learning Machines For Nonstationarity Detection Using Surrogates **3**

Pierre Borgnat, Patrick Flandrin, Cédric Richard, André Ferrari, and Hassan Amoud, Paul Honeine

1.1 Introduction 3
1.2 Revisiting stationarity 4
1.2.1 A time-frequency perspective 4
1.2.2 Stationarization via surrogates 5
1.3 Time-frequency learning machines 8
1.3.1 Reproducing kernels 8
1.3.2 The kernel trick, the representer theorem 9
1.3.3 Time-frequency learning machines: general principles . 9
1.3.4 Wigner distribution vs. spectrogram 11
1.4 A non-supervised classification approach 12
1.4.1 An overview on one-class classification 12
1.4.2 One-class SVM for testing stationarity 13
1.4.3 Spherical multidimensional scaling 15
1.5 Illustration 15
1.6 Conclusion 16

Bibliography **19**

Part I

This is a Part

Chapter 1

Time-Frequency Learning Machines For Nonstationarity Detection Using Surrogates

Pierre Borgnat, Patrick Flandrin

Laboratoire de Physique, UMR CNRS 5672, École Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon, France

Cédric Richard, André Ferrari

Laboratoire Fizeau, UMR CNRS 6525 Observatoire de la Côte d'Azur, Université de Nice Sophia-Antipolis, France

Hassan Amoud, Paul Honeine

Institut Charles Delaunay, FRE CNRS 2848, Université de technologie de Troyes, 12 rue Marie Curie, 10010 Troyes, France

1.1 Introduction 3
1.2 Revisiting stationarity 4
1.2.1 A time-frequency perspective 4
1.2.2 Stationarization via surrogates 5
1.3 Time-frequency learning machines 8
1.3.1 Reproducing kernels 8
1.3.2 The kernel trick, the representer theorem 9
1.3.3 Time-frequency learning machines: general principles 9
1.3.4 Wigner distribution vs. spectrogram 11
1.4 A non-supervised classification approach 12
1.4.1 An overview on one-class classification 12
1.4.2 One-class SVM for testing stationarity 13
1.4.3 Spherical multidimensional scaling 15
1.5 Illustration 15
1.6 Conclusion 16

1.1 Introduction

Time-frequency representations provide a powerful tool for nonstationary signal analysis and classification, supporting a wide range of applications [12]. As opposed to conventional Fourier analysis, these techniques reveal the evolution in time of the spectral content of signals. In [7, 39], time-frequency

analysis is used to test stationarity of any signal. The proposed method consists of a comparison between global and local time-frequency features. The originality is to make use of a family of stationary surrogate signals for defining the null hypothesis of stationarity and, based upon this information, to derive statistical tests. An open question remains, however, about how to choose relevant time-frequency features.

Over the last decade, a number of new pattern recognition methods based on reproducing kernels have been introduced. These learning machines have gained popularity due to their conceptual simplicity and their outstanding performance [30]. Initiated by Vapnik's Support Vector Machines (SVM) [36], they offer now a wide class of supervised and unsupervised learning algorithms. In [17, 18, 19], the authors have shown how the most effective and innovative learning machines can be tuned to operate in the time-frequency domain. The present paper follows this line of research by taking advantage of learning machines to test and quantify stationarity. Based on one-class support vector machines, our approach uses the entire time-frequency representation and does not require arbitrary feature extraction. Applied to a set of surrogates, it provides the domain boundary that includes most of these stationarized signals. This allows us to test the stationarity of the signal under investigation.

This paper is organized as follows. In Section 1.2, we introduce the surrogate data method to generate stationarized signals, namely, the null hypothesis of stationarity. The concept of time-frequency learning machines is presented in Section 1.3, and applied to one-class SVM in order to derive a stationarity test in Section 1.4. The relevance of the latter is illustrated by simulation results in Section 1.5.

1.2 Revisiting stationarity

1.2.1 A time-frequency perspective

Harmonizable processes define a general class of nonstationary processes whose spectral properties, which are potentially time-dependent, can be revealed by suitably chosen time-varying spectra. This can be achieved, e.g., with the Wigner-Ville Spectrum (WVS) [12], defined as

$$W_x(t,f) := \int \mathrm{E}\left\{x\left(t+\tau/2\right)x^*\left(t-\tau/2\right)\right\} e^{-i2\pi f\tau}\, d\tau, \tag{1.1}$$

where x stands for the analyzed process. Such a definition guarantees furthermore that second order stationary processes, which are a special case of harmonizable processes, have a time-varying spectrum that simply reduces to

the classical (stationary, time-independent) Power Spectrum Density (PSD) at every time instant.

In practice, the WVS has to be estimated on the basis of a single observed realization, a standard procedure amounting to make use of spectrograms (or multitaper variations [4]) defined as

$$S_x(t,f;h) := \left| \int x(\tau)\, h^*(\tau - t)\, e^{-i2\pi f\tau}\, d\tau \right|^2, \tag{1.2}$$

where h stands for some short-time observation window. In this case too, the concept of stationarity still implies time-independence, the time-varying spectra identifying, at each time instant, to some frequency smoothed version of the PSD. It follows however from this TF interpretation that, from an operational point of view, stationarity cannot be an *absolute* property. A more meaningful approach is to switch to a notion of *r*elative stationarity to be understood as follows: when considered over a given observed time scale, a process will be referred to as *stationary relative to this observation scale* if its time-varying spectrum undergoes no evolution or, in other words, if the local spectra $S_x(t_n, f; h)$ at all different time instants $\{t_n; n = 1, \ldots, N\}$ are statistically similar to the global (average) spectrum

$$\bar{S}_x(t_n, f; h) := \frac{1}{N} \sum_{n=1}^{N} S_x(t_n, f; h) \tag{1.3}$$

obtained by marginalization.

Based on this key point, one can imagine to design stationarity tests via some comparison between local and global features within a given observation scale [7] or, more generally, to decide whether an actual observation differs significantly from a stationary one within this time span. The question is therefore to have access to some stationary reference that would share with the observation the same global frequency behavior, while having a time-varying spectrum constant over time. As proposed in [7], an answer to this question can be given by the introduction of surrogate data.

1.2.2 Stationarization via surrogates

The general idea is to build a reference of stationarity directly from the signal itself, by generating a family of stationarized signals which have the same density spectrum as the initial signal. Indeed, given a density spectrum, non-stationary signals differ from stationary ones by temporal structures encoded in the spectrum phase. The surrogate data technique [35] is an appropriate solution to generate a family of stationarized signals, by keeping unchanged the magnitude of the Fourier transform $X(f)$ of the initial signal $x(t)$, and replacing its phase by an i.i.d. one. Each surrogate signal $x_\ell(t)$ results from the inverse Fourier transform of the modified spectrum, namely,

$$X_\ell(f) = |X(f)|\, e^{j\phi_\ell(f)},$$

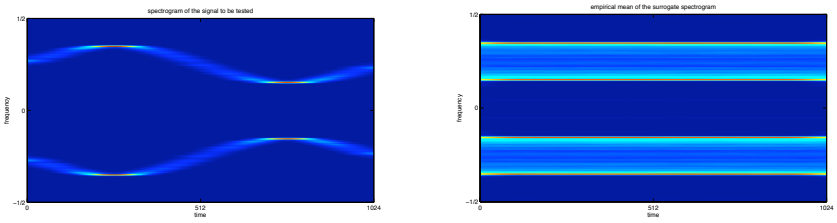

FIGURE 1.1: Spectrogram of a FM signal (left) and empirical mean of the spectrograms of its surrogates (right).

with $\phi_\ell(f)$ drawn from the uniform distribution over the interval $[-\pi, \pi[$. This leads to as many stationary surrogate signals, $x_1, \ldots, x_n$, as phase randomizations $\phi_1(f), \ldots, \phi_n(f)$ are operated. An illustration of the effectiveness of this approach in terms of its TF interpretation is given in Figure 1.1.

It has been first proved in [7] that surrogates are wide-sense stationary, i.e., their first and second order moments are time-shift invariant. More recently, it has been established in [26] that surrogates are strict-sense stationary, the proof proceeding as follows. Let us derive the invariance with respect to time shifts of the $(L+1)$-th order cumulant of the surrogate signal $x(t)$, where the subscript ℓ has been dropped for clarity

$$c(t; t_1, \ldots, t_L) = \mathrm{cum}(x^{\epsilon_0}(t), x^{\epsilon_1}(t+t_1), \ldots, x^{\epsilon_L}(t+t_L))$$

where $\epsilon_i = \pm 1$ and $x^{\epsilon_i}(t) = x^*(t)$ when $\epsilon_i = -1$ (we suggest the reader to refer, e.g., [1], for a detailed description of the tools related to high-order analysis of complex random processes). Let $\Phi(u) = \mathrm{E}[e^{j\phi u}]$ be the characteristic function of the random phase ϕ. As it is uniformly distributed over $[-\pi, \pi[$, note that

$$\Phi(k) = 0, \forall k \in \mathcal{Z}^*. \tag{1.4}$$

Using the multilinearity of the cumulants, we have

$$c(t; t_1, \ldots, t_L) = \int |X(f_0)| \cdots |X(f_L)|\, C(f_0, \ldots, f_L)\, e^{j2\pi t \sum_{i=0}^{L} \epsilon_i f_i}\, e^{j2\pi \sum_{i=1}^{L} \epsilon_i t_i f_i}\, df_0 \cdots df_L$$

where $C(f_0, \ldots, f_L) = \mathrm{cum}(e^{j\epsilon_0 \phi(f_0)}, \ldots, e^{j\epsilon_L \phi(f_L)})$. If one variable f_i is different from the others, the corresponding random variable $e^{j\epsilon_0 \phi(f_i)}$ is independent from the others and $C(f_0, \ldots, f_L) = 0$. Consequently, the joint cumulant of the surrogate simplifies to

$$c(t; t_1, \ldots, t_L) = C_{L+1} \int |X(f)|^{L+1}\, e^{j2\pi f t \sum_{i=0}^{L} \epsilon_i}\, e^{j2\pi f \sum_{i=1}^{L} \epsilon_i t_i} df$$

where $C_{L+1} = \mathrm{cum}(e^{j\epsilon_0\phi}, \ldots, e^{j\epsilon_L\phi})$. Application of the Leonov-Shiryaev formula to this cumulant leads to

$$C_{L+1} = \sum_{\mathcal{P}} (|\mathcal{P}| - 1)!\,(-1)^{|\mathcal{P}|-1} \prod_{B\in\mathcal{P}} \Phi(\textstyle\sum_{i\in B} \epsilon_i) \tag{1.5}$$

where $\mathcal{P}$ runs through the list of all the partitions of $\{0, \ldots, L\}$ and B runs through the list of all the blocks of the partition $\mathcal{P}$. This expression can be simplified using (1.4) and noting that $\sum_{i\in B} \epsilon_i \in \mathcal{Z}$. Consequently, $\Phi(\sum_{i\in B} \epsilon_i)$ is non-zero, and necessarily equal to 1, if and only if $\sum_{i\in B} \epsilon_i = 0$.

- If L is even, whatever $\mathcal{P}$, at least one block B of $\mathcal{P}$ has an odd cardinal. For this block, we have $\sum_{i\in B} \epsilon_i \in \mathcal{Z}^*$ and, consequently, $C_{L+1} = 0$.
- If L is odd, the product in (1.5) is non-zero, and thus equal to 1, if and only if $\sum_{i\in B} \epsilon_i = 0$ for all B of $\mathcal{P}$. Since $\sum_B \sum_{i\in B} \epsilon_i = \sum_{i=0}^{L} \epsilon_i$, this product is non-zero if, and only if, $\sum_{i=0}^{L} \epsilon_i = 0$.

As a conclusion, high-order cumulants of the surrogate signal $x(t)$ are non-zero only if $\sum_{i=0}^{L} \epsilon_i = 0$. This implies that $x(t)$ is a circular complex random signal. Moreover, substitution of this constraint in (1.2.2) leads to

$$c(t; t_1, \ldots, t_L) = C_{L+1} \int A(f)^{L+1}\, e^{j2\pi f \sum_{i=1}^{L} \epsilon_i t_i}\, df \tag{1.6}$$

which proves that surrogates are strict-sense stationary.

Remark — Making use of strictly stationary surrogates proved effective for detecting nonstationarities in various scenarii, but the tests happen to be very sensitive. For instance, when applied to realizations of actual stationary processes, e.g., AR, the rejection rate of the null hypothesis turns out to be higher than the prescribed confidence level [7, 38]. In a related way, one key point of the approach is to encompass in a common (time-frequency) framework stochastic and deterministic situations, stationarity referring to pure tones in the latter case. In this case too, surrogates cannot really reproduce the supposed stationarity of the observation. This is a natural outcome of the intrinsically stochastic generation of surrogates, but this makes again the test somehow pessimistic. The observation of such remaining limitations in the use of classical surrogates for testing stationarity prompts to think about related, possibly more versatile constructions. One possibility in this direction is, rather than *replacing* the spectrum phase by an i.i.d. sequence, to *modify* the original phase by adding some random phase noise to it. Depending on the nature and the level of this added phase noise, one can get this way a controlled transition from the original process (be it stationary or not) to its stationary counterpart [6].

Once a collection of stationarized surrogate signals has been synthesized, different possibilities are offered to test the initial signal stationarity [7, 39].

A potential approach is to extract some features from the surrogate signals such as distance between local and global spectra, and to characterize the null hypothesis of stationarity by the statistical distribution of their variations in time. Another approach is based on statistical pattern recognition. It consists of considering surrogate signals as a learning set, and using it to estimate the support of the distribution of the stationarized signals. This will be detailed further in the next section.

1.3 Time-frequency learning machines

Most pattern recognition algorithms can be expressed in terms of inner products only, involving pairs of input data. Replacing these inner products with a (reproducing) kernel provides an efficient way to implicitly map the data into a high-dimensional space, and apply the original algorithm in this space. Calculations are then carried out without making direct reference to the nonlinear mapping applied to input data. This so-called *kernel trick* is the main idea behind (kernel) learning machines. In this section, we show that learning machines can be tuned to operate in the time-frequency domain by a proper choice of kernel. Refer to [18] for more details.

1.3.1 Reproducing kernels

Let $\mathcal{X}$ be a subspace of $\mathcal{L}_2(\mathcal{C})$, the space of finite-energy complex signals, equipped with the usual inner product defined by $\langle x_i, x_j \rangle = \int_t x_i(t)\, x_j^*(t)\, dt$ and its corresponding norm. A kernel is a function $\kappa(x_i, x_j)$ from $\mathcal{X} \times \mathcal{X}$ to $\mathcal{C}$, with hermitian symmetry. It is said to be *positive definite* on $\mathcal{X}$ if [2]

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i\, a_j^*\, \kappa(x_i, x_j) \geq 0 \tag{1.7}$$

for all $n \in \mathcal{N}$, $x_1, \ldots, x_n \in \mathcal{X}$ and $a_1, \ldots, a_n \in \mathcal{C}$. It can be shown that every positive definite kernel κ is the reproducing kernel of a Hilbert space $\mathcal{H}$ of functions from $\mathcal{X}$ to $\mathcal{C}$, that is,

1. the function $\kappa_{x_j} : x_i \mapsto \kappa_{x_j}(x_i) = \kappa(x_i, x_j)$ belongs to $\mathcal{H}$, for all $x_j \in \mathcal{X}$;
2. one has $\Theta(x_j) = \langle \Theta, \kappa_{x_j} \rangle_{\mathcal{H}}$ for all $x_j \in \mathcal{X}$ and $\Theta \in \mathcal{H}$,

where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product in $\mathcal{H}$. It suffices to consider the subspace $\mathcal{H}_0$ induced by the functions $\{\kappa_x\}_{x \in \mathcal{X}}$, and equip it with the following inner product

$$\langle \Theta_1, \Theta_2 \rangle_{\mathcal{H}_0} = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{i,1}\, a_{i,2}^*\, \kappa(x_i, x_j), \tag{1.8}$$

where $\Theta_1 = \sum_{i=1}^{n} a_{i,1}\,\kappa_{x_i}$ and $\Theta_2 = \sum_{i=1}^{m} a_{i,2}\,\kappa_{x_i}$ are elements of $\mathcal{H}_0$. We fill this incomplete Hilbertian space according to [2], so that every Cauchy sequence converges in that space. Thus, we obtain the Hilbert space $\mathcal{H}$ induced by the reproducing kernel κ, called a *reproducing kernel Hilbert space* (RKHS). One can show that every reproducing kernel is positive definite [2]. An example of kernel is the Gaussian kernel defined by $\kappa(x_i, x_j) = \exp(-\|x_i - x_j\|^2/2\sigma^2)$, with σ the kernel bandwidth. Other examples of reproducing kernels, and rules for designing and combining them, can be found, e.g., in [16, 36].

1.3.2 The kernel trick, the representer theorem

Substituting Θ by κ_{x_i} in item 2 of the definition of RKHS in Section 1.3.1, we get the following fundamental property

$$\kappa(x_i, x_j) = \langle \kappa_{x_i}, \kappa_{x_j} \rangle_{\mathcal{H}} \tag{1.9}$$

for all $x_i, x_j \in \mathcal{X}$. Therefore, $\kappa(x_i, x_j)$ gives the inner product in $\mathcal{H}$, the so-called *feature space*, of the images κ_{x_i} and κ_{x_j} of any pair of input data x_i and x_j, without having to evaluate them explicitly. This principle is called the *kernel trick*. It can be used to transform any linear data processing technique into a non-linear one, on the condition that the algorithm can be expressed in terms of inner products only, involving pairs of the input data. This is achieved by substituting each inner product $\langle x_i, x_j \rangle$ by a non-linear kernel $\kappa(x_i, x_j)$, leaving the algorithm unchanged and incurring essentially the same computational cost. In conjunction with the kernel trick, the representer theorem is a solid foundation of kernel learning machines such as SVM [28]. This theorem states that any function Θ of $\mathcal{H}$ minimizing a regularized criterion of the form

$$J((x_1, y_1, \Theta(x_1)), \ldots, (x_n, y_n, \Theta(x_n))) + \rho(\|\Theta\|^2_{\mathcal{H}}), \tag{1.10}$$

with ρ a strictly monotonic increasing function on $\mathcal{R}_+$, can be written as a kernel expansion in terms of the available data, namely,

$$\Theta(\cdot) = \sum_{j=1}^{n} a_j\,\kappa(\cdot, x_j). \tag{1.11}$$

In order to prove this, note that any function Θ of the space $\mathcal{H}$ can be decomposed as $\Theta(\cdot) = \sum_{j=1}^{n} a_j\,\kappa(\cdot, x_j) + \Theta_\perp(\cdot)$, where $\langle \Theta_\perp(\cdot), \kappa(\cdot, x_j) \rangle_{\mathcal{H}} = 0$ for all $j = 1, \ldots, n$. Using this with equation (1.9), we see that $\Theta_\perp$ does not affect the value of $\Theta(x_i)$, for all $i = 1, \ldots, n$. Moreover, we verify that (1.11) minimizes ρ since $\rho(\|\sum_{j=1}^{n} a_j\,\kappa(\cdot, x_j)\|^2_{\mathcal{H}} + \|\Theta_\perp\|^2_{\mathcal{H}}) \geq \rho(\|\sum_{j=1}^{n} a_j\,\kappa(\cdot, x_j)\|^2_{\mathcal{H}})$. This is the essence of the representer theorem.

1.3.3 Time-frequency learning machines: general principles

In this section, we investigate the use of kernel learning machines for pattern recognition in the time-frequency domain. To clarify the discussion, we

shall first focus on the Wigner distribution. This will be followed by an extension to other time-frequency distributions, linear and quadratic. Below, $\mathcal{A}_n$ denotes a training set containing n instances $x_i \in \mathcal{X}$ and the desired outputs or labels $y_i \in \mathcal{Y}$.

Among the myriad of time-frequency representations that have been proposed, the Wigner distribution is considered fundamental in a number of ways. Its usefulness derives from the fact that it satisfies many desired mathematical properties such as the correct marginal conditions and the weak correct-support conditions. This distribution is also a suitable candidate for time-frequency-based detection since it is covariant to time shifts and frequency shifts and it satisfies the unitarity condition [12]. The Wigner distribution is given by

$$W_x(t,f) := \int x(t+\tau/2)\, x^*(t-\tau/2)\, e^{-2j\pi f\tau}\, d\tau \tag{1.12}$$

where x is the finite energy signal to be analyzed (one can remark that, under mild conditions, the Wigner-Ville Spectrum that has been previously considered, see eq. (1.1), is nothing but the ensemble average of the Wigner distribution (1.12)). By applying conventional linear pattern recognition algorithms directly to time-frequency representations, we seek to determine a time-frequency pattern $\Phi(t,f)$ so that

$$\Theta(x) = \langle W_x, \Phi \rangle = \iint W_x(t,f)\, \Phi(t,f)\, dt\, df \tag{1.13}$$

optimizes a given criterion J of the general form (1.10). The principal difficulty encountered in solving such problems is that they are typically very high dimensional, the size of the Wigner distributions calculated from the training set being quadratic in the length of signals. This makes pattern recognition based on time-frequency representations time-consuming, if not impossible, even for reasonably-sized signals. With the kernel trick and the representer theorem, kernel learning machines eliminate this computational burden. It suffices to consider the following kernel

$$\kappa_W(x_i, x_j) = \langle W_{x_i}, W_{x_j} \rangle, \tag{1.14}$$

and note that W_{x_i} and W_{x_j} do not need to be computed since, by the unitarity of the Wigner distribution, we have

$$\kappa_W(x_i, x_j) = |\langle x_i, x_j \rangle|^2. \tag{1.15}$$

We verify that κ_W is a positive definite kernel by writing condition (1.7) as $\| \sum_j a_j\, W_{x_j} \|^2 \geq 0$, which is clearly verified. We are now in a position to construct the RKHS induced by this kernel, and denoted by $\mathcal{H}_W$. It is obtained by completing the space $\mathcal{H}_0$ defined below with the limit of every Cauchy sequence

$$\mathcal{H}_0 = \{\Theta : \mathcal{X} \to \mathcal{R} \mid \Theta(\cdot) = \sum_j a_j\, |\langle \cdot\,, x_j \rangle|^2, a_j \in \mathcal{R}, x_j \in \mathcal{X}\}. \tag{1.16}$$

Thus, we can use the kernel (1.15) with any kernel learning machine proposed in the literature to perform pattern recognition tasks in the time-frequency domain. Thanks to the representer theorem, solution (1.11) allows for a time-frequency distribution interpretation, $\Theta(x) = \langle W_x, \Phi_W \rangle$, with

$$\Phi_W = \sum_{j=1}^{n} a_j \, W_{x_j}. \tag{1.17}$$

This equation is directly obtained by combining (1.11) and (1.13). We should again emphasize that the coefficients a_j are estimated without calculating any Wigner distribution. The time-frequency pattern Φ_W can be subsequently evaluated with (1.17), in an iterative manner, without suffering the drawback of storing and manipulating a large collection of Wigner distributions. The inherent sparsity of the coefficients a_j produced by most of the kernel learning machines, a typical example of which is the SVM algorithm, may speed-up the calculation of Φ_W.

1.3.4 Wigner distribution vs. spectrogram

Let $R_x(t,f)$ be a given time-frequency representation of a signal x. Proceeding as in the previous section with the Wigner distribution, we are led to optimization problems that only involve inner products between time-frequency representations of training signals:

$$\kappa_R(x_i, x_j) = \iint R_{x_i}(t,f) \, R_{x_j}(t,f) \, dt \, df = \langle R_{x_i}, R_{x_j} \rangle. \tag{1.18}$$

This can offer significant computational advantages. A well-known time-frequency representation is the spectrogram (1.2), whose definition can be recast as

$$S_x(t,f;h) = \left| \langle x, h_{t,f} \rangle \right|^2,$$

with $h_{t,f}(\tau) := h(\tau - t) \, e^{2j\pi f \tau}$. The inner product between two spectrograms, say S_{x_i} and S_{x_j}, is given by the kernel [18]

$$\kappa_S(x_i, x_j) = \iint |\langle x_i, h_{t,f} \rangle \langle x_j, h_{t,f} \rangle|^2 \, dt \, df.$$

Computing this kernel for any pair of surrogate signals yields

$$\kappa_S(x_i, x_j) = \iint \left| \langle |X| e^{j\phi_i}, H_{t,f} \rangle \, \langle |X| e^{j\phi_j}, H_{t,f} \rangle \right|^2 dt \, df, \tag{1.19}$$

where $H_{t,f}$ is the Fourier transform of $h_{t,f}$. This has to be contrasted with the Wigner distribution which, with its unitarity property, leads to some substantial computational reduction since

$$\kappa_W(x_i, x_j) = |\langle x_i, x_j \rangle|^2 = \left| \int |X(f)|^2 \, e^{j(\phi_i(f) - \phi_j(f))} \, df \right|^2. \tag{1.20}$$

We emphasize here that there is no need to compute and manipulate the surrogates and their time-frequency representations. Given $|X(f)|$, only the random phases $\phi_i(f)$ and $\phi_j(f)$ are required to evaluate the kernel κ_W of eq. (1.20).

For the sake of simplicity, we illustrated this section with the spectrogram. However, kernel learning machines can use any time-frequency kernels to perform pattern recognition tasks in the time-frequency domain, as extensively studied in [17, 18, 19]. In the next section, we present the one-class SVM problem to test stationarity using surrogate signals.

1.4 A non-supervised classification approach

Adopting a viewpoint rooted in statistical learning theory by considering the collection of surrogate signals as a learning set, and using it to estimate the support of the distribution of stationarized data, avoids solving the difficult problem of density estimation that would be a pre-requisite in parametric methods. Let us make this approach, which consists of estimating quantiles of multivariate distributions, more precise.

1.4.1 An overview on one-class classification

In the context considered here, the classification task is fundamentally a one-class classification problem and differs from conventional two-class pattern recognition problems in the way how the classifier is trained. The latter uses only target data to perform outlier detection. This is often accomplished by estimating the probability density function of the target data, e.g., using a Parzen density estimator [24]. Density estimation methods however require huge amounts of data, especially in high dimensional spaces, which makes their use impractical. Boundary-based approaches attempt to estimate the quantile function defined by $Q(\alpha) := \inf\{\lambda(\mathcal{S}) : P(\mathcal{S}) := \int_{\omega\in\mathcal{S}} \mu(d\omega) \geq \alpha\}$ with $0 < \alpha \leq 1$, where $\mathcal{S}$ denotes a subset of the signal space $\mathcal{X}$ that is measurable with respect to the probability measure μ, and $\lambda(\mathcal{S})$ its volume. Estimators that reach this infinimum, in the case where P is the empirical distribution, are called minimum volume estimators. The first boundary-based approach was probably introduced in [27], where the authors consider a class of closed convex boundaries in $\mathcal{R}^2$. More sophisticated methods were described in [21, 22]. Nevertheless, they are based upon neural networks training and therefore suffer from the same drawbacks such as slow convergence and local minima. Inspired by support vector machines, the support vector data description algorithm proposed in [34] encloses data in a minimum volume hypersphere. More flexible boundaries can be obtained by using kernel functions, that map the data into a high-dimensional feature space. In the case

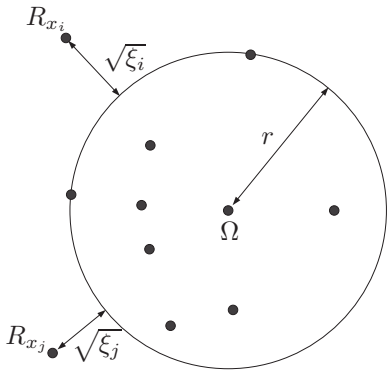

FIGURE 1.2: Support vector data description algorithm

of normalized kernel functions, that is, such that $\kappa(x, x) = 1$ for all x, this approach is equivalent to the one-class support vector machines introduced in [29], which use a maximum margin hyperplane to separate data from the origin. The generalization performance of these algorithms were investigated in [29, 31, 37] via the derivation of bounds. In what follows, we shall focus on the support vector data description algorithm.

1.4.2 One-class SVM for testing stationarity

Inspired by SVM for classification, the one-class SVM allows the description of the density distribution of a single class [33]. The main purpose is to enclose the training data into a minimum volume hypersphere, thus defining a domain boundary. Any data outside this volume may be considered as an outlier, and its distance to the center of the hypersphere allows a measure of its novelty. Here, we propose to use the set of surrogate signals to derive the hypersphere of stationarity, in the time-frequency domain defined by a reproducing kernel as given in Section 1.3.

Consider a set of n surrogate signals, $x_1, \ldots, x_n$, computed from a given signal x. Let $R_{x_1}, \ldots, R_{x_n}$ denote their time-frequency representations and κ the corresponding reproducing kernel. In this domain, we seek the hypersphere that contains most of these representations. Its center Ω and radius r are obtained by solving the optimization problem

$$\begin{aligned} &\min_{\Omega, r, \xi} \quad r^2 + \frac{1}{n\nu} \sum_{i=1}^{n} \xi_i \\ &\text{subject to} \quad \|R_{x_i} - \Omega\|^2 \leq r^2 + \xi_i, \quad \xi_i \geq 0, \quad i = 1, \ldots, n. \end{aligned}$$

As illustrated in Fig. 1.2, parameter $\nu \in]0, 1]$ controls the tradeoff between the radius r to be minimized, and the number of training data outside the hypersphere characterized by the slack variables $\xi_i = (\|R_{x_i} - \Omega\|^2 - r^2)_+$.

Using Lagrangian principle, the optimization problem is reduced to

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{n} \alpha_i \, \kappa(x_i, x_i) - \sum_{i,j=1}^{n} \alpha_i \, \alpha_j \, \kappa(x_i, x_j) \\ \text{subject to} \quad & \sum_{i=1}^{n} \alpha_i = 1, \quad 0 \leq \alpha_i \leq \frac{1}{n\nu}, \quad i = 1, \ldots, n, \end{aligned} \tag{1.21}$$

which can be solved with quadratic programming techniques. The resulting non-zero Lagrange multipliers α_i yield the center $\Omega = \sum_i \alpha_i \, R_{x_i}$, and the radius $r = \|R_{x_\ell} - \Omega\|$ with x_ℓ any data having $0 < \alpha_\ell < \frac{1}{n\nu}$.

For any signal x, the (squared) distance of its time-frequency representation to the center Ω can be written as

$$\|R_x - \Omega\|^2 = \kappa(x, x) - 2 \sum_{i=1}^{n} \alpha_i \, \kappa(x, x_i) + \sum_{i,j=1}^{n} \alpha_i \, \alpha_j \, \kappa(x_i, x_j).$$

As explained previously, we do not need to compute time-frequency representations to calculate this score, since only the values of the kernel are required. The coefficients α_i are obtained by solving (1.21), requiring only the evaluation of κ for training data. This *kernel trick* is also involved in the proposed decision function, defined by comparing the test statistics $\Theta(x) = \|R_x - \Omega\|^2 - r^2$ to a threshold γ:

$$\Theta(x) \underset{\text{stat.}}{\overset{\text{nonstat.}}{\gtrless}} \gamma. \tag{1.22}$$

The signal x under study is considered as nonstationary if its time-frequency representation lies outside the hypersphere of squared radius $r^2 + \gamma$; otherwise, it is considered as stationary. The threshold γ has a direct influence upon the test performance [30]. For instance, with a probability greater than $1 - \delta$, one can bound the probability of false positive by

$$\Delta = \frac{1}{\gamma n} \sum_{i=1}^{n} \xi_i + \frac{6\omega^2}{\gamma\sqrt{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}, \tag{1.23}$$

where ω is the radius of the ball centered at the origin containing the support of the probability density function of the class of surrogate signals. Here γ can be fixed arbitrarily in eq. (1.22), so as to set the required false positive probability, for which Δ is an upper bound.

We shall now propose another use of eq. (1.23) as a measure of stationarity of the signal x under investigation. If x lies inside the hypersphere of the surrogate class, the score of stationarity is arbitrarily fixed to one. Else, one can set γ depending on the signal x to $\|R_x - \Omega\|^2 - r^2$, so that the signal would lie on the decision boundary. Then, eq. (1.23) gives a bound $\Delta(x)$ on the false positive probability that should be assumed for the signal to be classified as stationary. The closer x is to the hypersphere boundary, the closer to one

$\Delta(x)$ is; the closer $\Delta(x)$ is to zero, the greater the contrast between x and the surrogates is. Hence, $\Delta(x)$ has the meaning of a stationarity score. See [7] for more details.

1.4.3 Spherical multidimensional scaling

Multidimensional scaling (MDS) is a classical tool in data analysis and visualization [9]. It aims at representing data in a d-dimensional space, where d is specified *a priori*, such that the resulting distances reflect in some sense the distances in the higher-dimensional space. The neighborhood between data is preserved, whereas dissimilar data tend to remain distant in the new space. MDS algorithm requires only the distances between data in order to embed them into the new space. Consider the set of time-frequency representations of surrogate signals $\{R_{x_1}, \ldots, R_{x_n}\}$, and the inner product between two representations defined as in (1.18). We can apply classical MDS in order to visualize the data in a low-dimensional Euclidean space. On the condition that R_x satisfies the global energy distribution property $\iint R_x(t,f)\,dt\,df \propto \int |x(t)|^2 dt$, such as the spectrogram or the Wigner distribution, the time-frequency representations of surrogate signals lie on an hypersphere centered at the origin. This non-Euclidean geometry makes it desirable to place restrictions on the configuration obtained from the MDS analysis. This can be done by using a Spherical MDS technique, as proposed in [9], or more recently in [25], which forces points to lie on the the two-dimensional surface of a sphere.

In the experimental results section, we shall use spherical MDS analysis to visualize the time-frequency configuration of the signal x under study and the surrogate signals, and the decision boundary that discriminates between the null hypothesis of stationarity and its nonstationary alternative.

1.5 Illustration

In order to test our method, we used the same two AM and FM signals as in [39]. While not covering all the situations of nonstationarity, these signals are believed to give meaningful examples. The AM signal is modeled as

$$x(t) = (1 + m\sin(2\pi t/T_0))\,e(t)$$

with $m \leq 1$, $e(t)$ a white Gaussian noise, T_0 the period of the AM. In the FM case,

$$x(t) = \sin(2\pi(f_0 t + m\sin(2\pi t/T_0)))$$

with f_0 the central frequency. Based on the relative values of T_0 and the signal duration T, three cases can be distinguished for each type, AM and FM:

- $T \gg T_0$: The signal contains a great number of oscillations. This periodicity indicates a stationary regime.
- $T \approx T_0$: Only one oscillation is available. The signal is considered as nonstationary.
- $T \ll T_0$: With a small portion of a period, there is no change in the amplitude or the frequency. It is considered as a stationary signal.

For each experiment reported in Fig. 1.3, 50 surrogate signals were generated from the AM or FM signal $x(t)$ to be tested. The results are displayed for $T_0 = T/100$, $T = T_0$ and $T_0 = 100\,T$, allowing to consider stationarity relatively to the ratio between observation time T and modulation period T_0. The one-class SVM algorithm was run with the spectrogram kernel (1.19) and parameter $\nu = 0.15$. Then, spherical MDS analysis was applied for visualization purpose only. In each figure, the surrogate signals are shown with blue stars and the signal to be tested with a red triangle. The minimum-volume domain of stationarity is represented by the black curve. It should be noticed that this curve and the data are projections from the high-dimensional space of time-frequency distributions onto a sphere in $\mathcal{R}^3$ for visualization, meaning that the representation is inherently distorted. The tested signals are clearly identified as nonstationary in the case $T = T_0$ (the red triangle of the signal being outside the circle corresponding to the minimum-volume domain of stationarity), and can be considered as stationary in the cases $T_0 = T/100$ and $T_0 = 100\,T$ (the red triangle of the signal being inside the circle). These results are consistent with those obtained in previous works, using either the distance or the time-frequency feature extraction approach [7]. Here, the test is performed without suffering from the prior knowledge required to extract relevant features.

1.6 Conclusion

In this paper, we showed how time-frequency kernel machines can be used to test stationarity. For any given signal, a set of stationarized surrogate signals is generated to train a one-class SVM, implicitly in the time-frequency domain by using an appropriate kernel. The originality here is the use of the whole time-frequency information, as opposed to feature extraction techniques where prior knowledge is required. This was proved effective for detecting nonstationarities with simulation results.

The resampling strategy actually used to generate surrogate signals is however quite strict in the sense that, after the phase replacement by some i.i.d. sequence, a possibly nonstationary signal is turned into a stationary one without any consideration about the fact that this property has to be understood in

a relative sense that incorporates the observation scale. In this respect, strict stationarity may appear as a too strong commitment, and prompts to think about more versatile constructions of stationary-relaxed surrogate signals. One perspective in this direction is to alter the original phase by embedding it into noise [6]. Depending on the nature and the level of these phase fluctuations, we could get this way a controlled transition from the original process to its stationary counterpart.

There has been preliminary attempts to use surrogates for testing the stationarity of some actual data: signals in biophysics [5], or in mechanics [8], and an adaptation to images is reported in [13].

Let us now turn to potential astronomical applications. Searching for evidence of non-stationarity in the temporal properties of astrophysical objects and phenomena are fundamental to their study, as in the case of Galactic black holes switching from one spectral state to another. For instance, among the basic properties characterizing an Active Galactic Nucleus, the X-ray variability is one of the most commonly used. This question has been explored by researchers in the case of NGC 4051 [15], Mrk 421 [10, 11], 3C 390.3 [14], and PKS 2155-304 [40]. One of the most popular approaches is to measure the fluctuation of the power spectral density (PSD), by fitting a simple power law model [20], or by estimating the so-called excess variance [23]. As an alternative, analysis in the time domain using the structure function (SF) is also often considered [32]. Note that PSD with model fitting and SF should provide equivalent information as they are related to the auto-correlation function of the process at hand. It is unfortunate that, with this techniques, it is not possible to prove non-stationarity in a model independent way. Nonlinear analysis using scaling index method, which measures both global and local properties of a phase-space portrait of time series, has also been considered in the literature [3].

Although its effective application to real astrophysical data has not yet been considered, it is believed that the approach proposed here could be a useful addition to such existing techniques by shedding a new light on stationarity vs. nonstationarity issues.

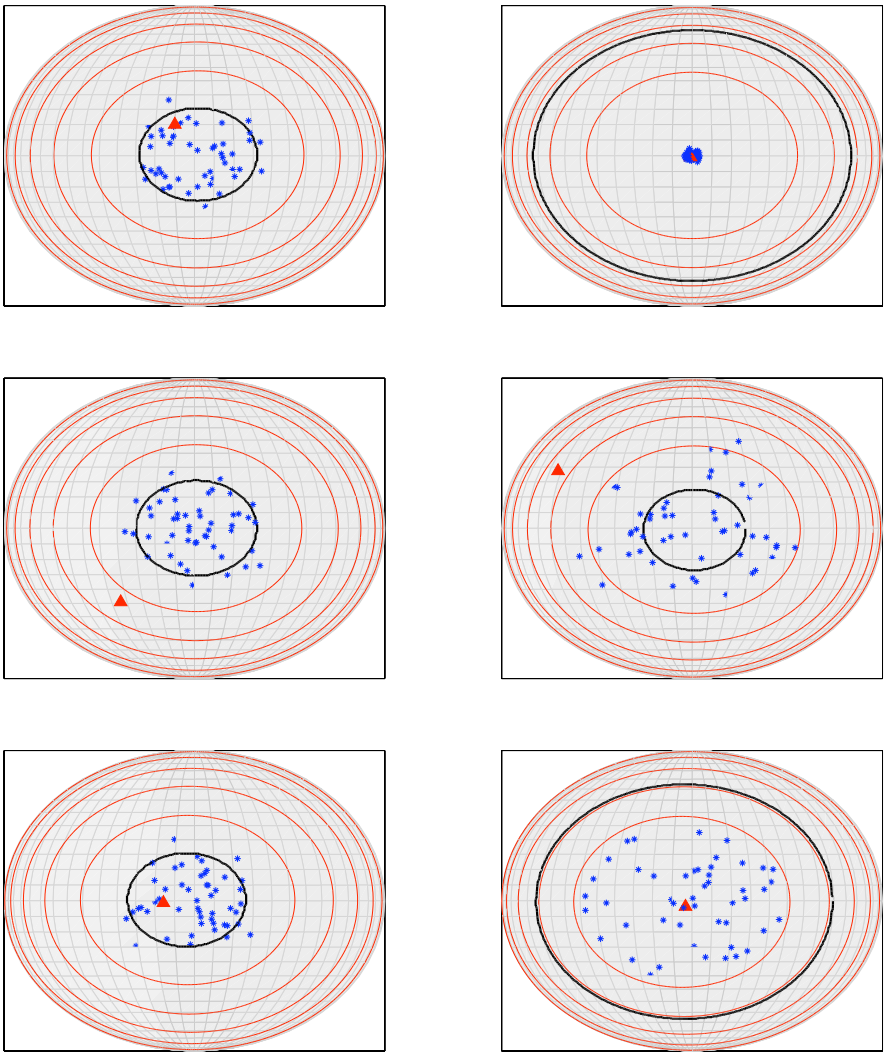

FIGURE 1.3: Spherical MDS representation of the surrogate signals (∗) and the test signal (▲), in AM (left) and FM (right) situations. From top to bottom, $T_0 = T/100$, $T = T_0$ and $T_0 = 100\,T$, with $T = 1024$. The minimum-volume domain of stationarity is represented by the black curve. The tested signals are identified as nonstationary in the case $T = T_0$ (the red triangle of the signal being outside the circle corresponding to the minimum-volume domain of stationarity), and can be considered as stationary in the cases $T_0 = T/100$ and $T_0 = 100\,T$ (the red triangle of the signal being inside the circle). Other parameters are as follows – (AM): $m = 0.5$; (FM): $f_0 = 0.25$, $m = 0.02$, and SNR = 10 dB.

Bibliography

[1] P. O. Amblard, M. Gaeta, and J. L. Lacoume. Statistics for complex variables and signals – part I and II. *Signal Processing*, 53(1):1 – 25, 1996.

[2] N. Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404, 1950.

[3] H. Atmanspacher, H. Scheingraber, and G. Wiedenmann. Determination of f(.) for a limited random point set. *Physical Review A*, 40(7):3954–3963, Oct 1989.

[4] M. Bayram and R.G. Baraniuk. Multiple window time-varying spectrum estimation. In W.J. Fitzgerald et al., editor, *Nonlinear and Nonstationary Signal Processing*. Cambridge Univ. Press, 2000.

[5] P. Borgnat and P. Flandrin. Stationarization via surrogates. *Journal of Statistical Mechanics: Theory and Experiment: Special issue UPoN 2008*, page P01001, January 2009.

[6] P. Borgnat, P. Flandrin, A. Ferrari, and C. Richard. Transitional surrogates. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Prague, Czech Republic, 2011. Accepted.

[7] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao. Testing stationarity with surrogates: A time-frequency approach. *IEEE Trans. Signal Processing*, 58(12):3459–3470, 2010.

[8] M. Chabert, B. Trajin, J. Regnier, and J. Faucher. Surrogate-based diagnosis of mechanical faults in induction motor from stator current measurements (regular paper). In *Condition Monitoring, Stratford upon Avon, 22/06/2010-24/06/2010*, page (electronic medium). The British Institute of Non Destructive Testing, 2010.

[9] T. F. Cox and M. A. A. Cox. Multidimensional scaling on a sphere. *Communications in Statistics: Theory and Methods*, 20:2943–2953, 1991.

[10] D. Emmanoulopoulos, I. M. McHardy, and P. Uttley. Variability studies in blazar jets with SF analysis: caveats and problems. In G. E. Romero, R. A. Sunyaev, & T. Belloni, editor, *IAU Symposium*, volume 275 of *IAU Symposium*, pages 140–144, February 2011.

[11] V. V. Fidelis and D. A. Iakubovskyi. The X-ray variability of Mrk 421. *Astronomy Reports*, 52:526–538, July 2008.

[12] P. Flandrin. *Time-Frequency/Time-Scale Analysis*. Academic Press Inc, 1998.

[13] P. Flandrin and P. Borgnat. Revisiting and testing stationarity. *J. Phys.: Conf. Series.*, 139:012004, 2008.

[14] M. Gliozzi, I. E. Papadakis, and C. Räth. Correlated spectral and temporal changes in 3C 390.3: a new link between AGN and Galactic black hole binaries? *Astronomy and Astrophysics*, 449:969–983, April 2006.

[15] A. R. Green, I. M. McHardy, and C. Done. The discovery of non-linear x-ray variability in ngc 4051. *Monthly Notices of the Royal Astronomical Society*, 305(2):309–318, 1999.

[16] R. Herbrich. *Learning kernel classifiers. Theory and algorithms*. The MIT Press, Cambridge, MA, 2002.

[17] P. Honeine and C. Richard. Signal-dependent time-frequency representations for classification using a radially gaussian kernel and the alignment criterion. In *Proc. IEEE Statistical Signal Processing*, pages 735–739, Madison, WI, USA, August 2007.

[18] P. Honeine, C. Richard, and P. Flandrin. Time-frequency learning machines. *IEEE Trans. Signal Processing*, 55:3930–3936, July 2007.

[19] P. Honeiné, C. Richard, P. Flandrin, and J.-B. Pothin. Optimal selection of time-frequency representations for signal classification: a kernel-target alignment approach. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Toulouse, France, 2006.

[20] A. Lawrence and I. Papadakis. X-ray variability of active galactic nuclei - A universal power spectrum with luminosity-dependent amplitude. *Astrophysical Journal*, 414:L85–L88, September 1993.

[21] M. Moya and D. Hush. Network contraints and multi-objective optimization for one-class classification. *Neural Networks*, 9(3):463–474, 1996.

[22] M. Moya, M. Koch, and L. Hostetler. One-class classifier networks for target recognition applications. In *Proc. World Congress on Neural Networks*, pages 797–801, 1993.

[23] K. Nandra, I. M. George, R. F. Mushotzky, T. J. Turner, and T. Yaqoob. Asca observations of seyfert 1 galaxies. i. data analysis, imaging, and timing. *The Astrophysical Journal*, 476(1):70, 1997.

[24] E. Parzen. On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33(3):1065–1076, 1962.

[25] R. Pless and I. Simon. Embedding images in non-flat spaces. In H. R. Arabnia and Y. Mun, editors, *Proc. International Conference on Imaging Science, Systems and Technology*, pages 182–188, Las Vegas, 2002.

[26] C. Richard, A. Ferrari, H. Amoud, P. Honeine, P. Flandrin, and P. Borgnat. Statistical hypothesis testing with time-frequency surrogates to check signal stationarity. In *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Dallas, USA, 2010.

[27] T. W. Sager. An iterative method for estimating a multivariate mode and isopleth. *Journal of the American Statistical Association*, 74(366):329–339, 1979.

[28] B. Schölkopf, R. Herbrich, and R. Williamson. A generalized representer theorem. Technical Report NC2-TR-2000-81, NeuroCOLT, Royal Holloway College, University of London, UK, 2000.

[29] B. Schölkopf and A.J. Smola. *Learning With Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, 2001.

[30] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[31] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[32] J. H. Simonetti, J. M. Cordes, and D. S. Heeschen. Flicker of extragalactic radio sources at two frequencies. *Astrophysical Journal*, 296:46–59, September 1985.

[33] D. M. J. Tax and R. P. W. Duin. Support vector domain description. *Pattern Recognition Letters*, 20(11-13):1191–1199, 1999.

[34] D. M. J. Tax and R. P. W. Duin. Support vector data description. *Machine Learning*, 54(1):45–66, 2004.

[35] J. Theiler, S. Eubank, A. Longtin, B. Galdrikian, and J. Doyne Farmer. Testing for nonlinearity in time series: the method of surrogate data. *Physica D: Nonlinear Phenomena*, 58(1-4):77–94, 1992.

[36] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[37] R. Vert. *Theoretical insights on density level set estimation, application to anomaly detection*. PhD thesis, Paris 11 - Paris Sud, 2006.

[38] J. Xiao, P. Borgnat, and P. Flandrin. Sur un test temps-fréquence de stationnarité. *Traitement du Signal*, 25(4):357–366, 2008. (in French, with extended English summary).

[39] J. Xiao, P. Borgnat, P. Flandrin, and C. Richard. Testing stationarity with surrogates - a one-class SVM approach. In *Proc. IEEE Statistical Signal Processing*, pages 720–724, Madison, WI, USA, August 2007.

[40] Y. H. Zhang, A. Treves, A. Celotti, Y. P. Qin, and J. M. Bai. Xmm-newton view of pks 2155–304: Characterizing the x-ray variability properties with epic pn. *The Astrophysical Journal*, 629(2):686, 2005.