Testing for high frequency features in a noisy signal

Mathieu Mezache, Marc Hoffmann, Human Rezaei, Marie Doumic

Testing for high frequency features in a noisy signal

Mathieu Mezache* Marc Hoffmann† Human Rezaei‡ Marie Doumic§

Abstract

The aim of this article is to detect high frequency (HF) features in a noisy signal. We propose a parametric characterization in the Fourier domain of the HF features. Then we introduce a procedure to evaluate these parameters and compute a p-value which assesses in a quantitative manner the presence or absence of such features, that we also call "oscillations". The procedure is well adapted for real 1-dimensional signals. If the signal analyzed has singular events in the low frequencies, the first step is a data-driven regularization of its Fourier transform. In the second step, the HF features parameters are estimated. The third step is the computation of the p-value thanks to a Monte Carlo procedure. The test is conducted on sanity-check signals where the ratio amplitude of the oscillations/level of the noise is entirely controlled. The test detects HF features even when the level of the noise is five times larger than the amplitude of the oscillations. The test is also conducted on signals from Prion disease experiments and confirms the presence of HF features in these signals.

Keywords: Discrete Fourier transform, hypothesis testing, spectral analysis, Monte Carlo methods, signal detection and filtering, Static Light Scattering, Prions.

Mathematical Subject Classification: 65T50, 62F03, 62M15, 65C05, 60G35, 92B15, 62P10.

* Sorbonne Universités, Inria, Université Paris-Diderot, CNRS, Laboratoire Jacques-Louis Lions, F-75005 Paris, France,mathieu.mezache@inria.fr

†Université Paris-Dauphine PSL, CEREMADE, Place du Maréchal de Lattre de Tassigny, F-75016 Paris, hoffmann@ceremade.dauphine.fr

‡INRA, UR892, Virologie Immunologie Moléculaires, 78350 Jouy-en-Josas, France, human.rezaei@inra.fr

§Sorbonne Universités, INRIA, Université Paris-Diderot, CNRS, Laboratoire Jacques-Louis Lions, F-75005 Paris, France, marie.doumic@inria.fr - Wolfgang Pauli Institute, c/o university of Vienna, Austria

Introduction

Motivation

In a one-dimensional signal, transient oscillations may reveal key features of the underlying processes. As an example, and original motivation for our study, fast oscillations have been visually observed in experimental measurements of the infectious agent in Prion diseases, see Figure 1.

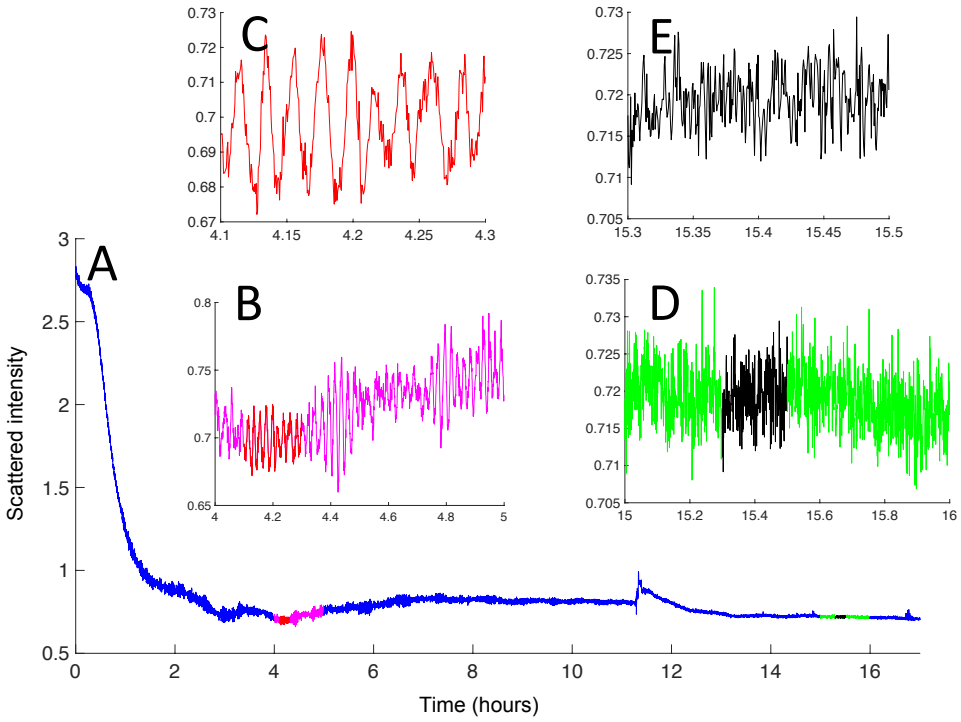

Figure 1: **Human PrP amyloid fibrils (Hu fibrils) depolymerisation monitored by Static Light Scattering** (see Appendix for details). A: The overall view of the $0.35\mu M$ Hu-fibrils depolymerisation at 55^0C. B-E correspond to a zoom-in on different time-segments of the depolymerisation curve A. As shown in B, from time 4h to time 5h oscillations have been observed when for time segment corresponding to time 15.3 to 15.5h only noise has been detected (D). (Figure taken from [9])

A major difficulty to infer such transient oscillations and to evaluate their significance is that they are mixed up with noise. Hence it is of major interest to rely on a rigorous procedure which detects high frequency (HF) features - amplitude, frequency - in real signals and to distinguish quantitatively these features of the signal from its noise.

To our knowledge, there exist only few methods to detect and estimate the HF features in a signal. The Singular Spectrum Analysis (SSA) introduced by Broomhead and Jones in [3] is one of those and allows one to visualize qualitative dynamics from noisy experimental data. The SSA is based on the decomposition of a time series or signal into several additive components interpreted as trend components, oscillatory components, and noise components. It was then widely used to identify intermittent or modulated oscillations in time series, see e.g. [22, 17, 10].

A statistical test of hypothesis to discriminate between potential oscillations and noise has been introduced in [1] and [16]. This test is called the Monte Carlo SSA and has been applied almost exclusively to meteorological data. Since SSA transforms the original data in a complex way, no theoritical result has yet been proved on the Monte Carlo SSA. Prior knowledge on the signal (such as the trend or assumptions on the noise) are also needed in order to calibrate the procedure and improve the result of the statistical test. The Monte Carlo SSA is by construction a non-parametric procedure and the oscillations detected by this test are not characterized quantitatively but qualitatively.

In this paper, we propose another method, based on the Fourier transform of the signal, to infer a parametric characterization of HF features, based on their amplitude and frequency detection. This method is detailed in Section 1. We then introduce a statistical test to discriminate HF features from noise in Section 3, apply our methodology to a simulated example in Section 2, and then to the experimental measurements of PrP protein displayed in Figure 12 in Section 4.

Model and assumptions

For some (large) $n \geq 1$, we have measurements y_i^n of a noisy signal localized around i/n, so that i is a location parameter and n a frequency parameter. We may idealise our data via a representation of the form

$$y_i^n = x_i^n + \sigma \xi_i^n, \qquad i = 0, \dots, n-1, \tag{1}$$

where $(x_i^n)_{0 \leq i \leq n-1}$ is the true (unknown) signal of interest and the ξ_i are independent and identically distributed noise measurement, that we assume here to be standard Gaussian. The quantity $\sigma > 0$ is a (fixed) noise level. In this nonparametric regression setting, we aim at detecting from the data $(y_i^n)_{0 \leq i \leq n-1}$ whether $(x_i^n)_{0 \leq i \leq n-1}$ exhibits *high-frequency features* (HF features) such as oscillations, a term that still needs to be defined properly. Since we do not know in advance whether such high-frequency features are present and where they are located, we need

to investigate the shape of $(x_i^n)_{0\leq i\leq n-1}$, which requires some smoothing in order to get rid of the noise $(\xi_i^n)_{0\leq i\leq n-1}$. However, any smoothing procedure tends to wipe out high-frequencies in the data, which is adversarial to our goal.

Results and organisation of the paper

The statistical test to differentiate HF features from noise in a signal is data-driven and is based on the study of the projection of the signal in the Fourier domain. We propose in Section 1 a parametric characterization of the HF features of a signal. This characterization also provides an algorithmic procedure for the computation of the HF features, implemented in the Python language at *https://github.com/mmezache/HFFTest* (see Appendix B). The procedure consists in three steps: in the first step, a regularization procedure is applied to the experimental data in order to smooth the fast variations that may exist in the low frequency range. The second step of the procedure is the detection and localization of significant peaks in the Fourier domain. The third step is the computation of the HF features parameters by selecting one of these peaks. The construction of the statistical test of hypothesis and the computation of the p-value is described in Section 2.

The numerical examples are performed in Section 3 with sanity-check signals. They are constructed around parameters which control their trend, their transient oscillations and their noise. We vary the ratio of the amplitude of the HF features over the noise level (i.e. its standard deviation), which sheds light on the robustness of the procedure: the transient oscillations are detected by the procedure even if the noise level is significantly high. The procedure is then applied to static light scattering (SLS) experiments of PrP^{Sc} fibrils, in Section 4. They are characterised by their singular slow-varying components (non-monotonous trend) and their fast-varying components (isolated discontinuous jumps, transient oscillations, noise). We compute the HF features parameters of SLS signal experiments for different initial concentration of PrP^{Sc}. We conclude that these signals have significant HF features, i.e. the signals display transient oscillations coming from biochemical reactions and not from the experimental noise.

1 Characterisation of high frequency features

The discrete Fourier transform (DFT) $\mathsf{DFT}_n : \mathbb{R}^n \to \mathbb{R}^n$ transfers a real-valued discrete signal $(x_i^n)_{0\leq i\leq n-1}$ of length n into a frequency domain via

$$\mathsf{DFT}_n\big[(x_i^n)_{0\leq i\leq n-1}\big] = \Big(\sum_{i=0}^{n-1} x_i^n \mathrm{e}^{-j2\pi ki/n}\Big)_{0\leq k\leq n-1} = \big(\vartheta_{n,k}\big)_{0\leq k\leq n-1}. \tag{2}$$

The single-sided amplitude spectrum gives all the information needed to visualise the signal $(x_i^n)_{0\leq i\leq n-1}$ in the Fourier basis.

Our typical experimental signals have a specific low frequency trend combined with HF features or transient oscillations that shall persist beyond denoising. The presence of a trend implies that there are large Fourier coefficients $\vartheta_{n,k}$ on the scale corresponding to the low frequency information. Transient oscillations can be characterised by large coefficients in mid or high frequencies that are relatively well localised. As displayed by the test signal in Figure 2, a typical signal displaying oscillations would thus consist, in the frequency domain, of large coefficients in the low frequency, then a decay to a minimum value, and then one or more peaks in mid or high frequencies and a decay as the frequency grows further. Hence HF features in a signal corresponds to a level of energy (measured by the norm of the DFT coefficients) at a specific distance from the low frequency DFT coefficients in the frequency domain (cf Figure 3).

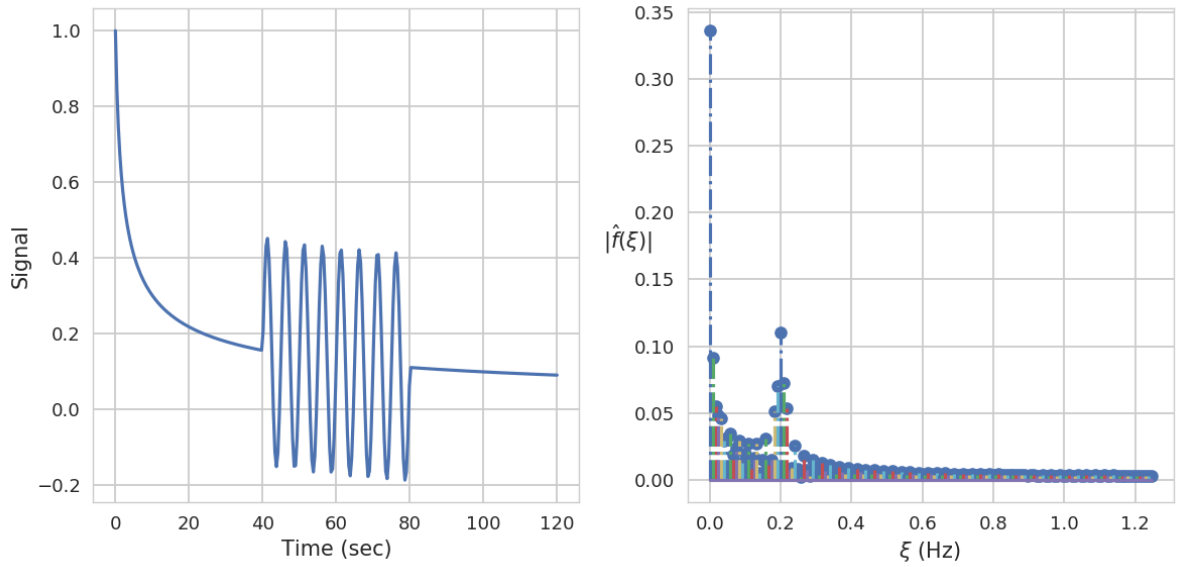

Figure 2: **Graph of a test signal with HF features and its single-sided amplitude spectrum. Left:** Plot of $z_i = f(0.4i)$ for $i = 0, \dots, 300$ where $f(x) = \frac{1}{\sqrt{x+1}} + 0.3 \sin\left(\frac{2\pi x}{5}\right) \mathbb{1}_{[40,80]}(x)$. **Right:** Plot of the amplitude spectrum of $(z_i)_{0 \le i \le 300}$.

For a discrete signal $(x_i^n)_{0 \le i \le n-1}$ given in terms of its Fourier transform $\vartheta_n = (\vartheta_{n,k})_{0 \le i \le n-1}$ via (2), we characterise a HF feature by two nonnegative parameters: a location parameter $\mathsf{g}(\vartheta_n)$ (in the frequency domain) and an intensity parameter $\mathsf{d}(\vartheta_n)$ (see Figure 3).

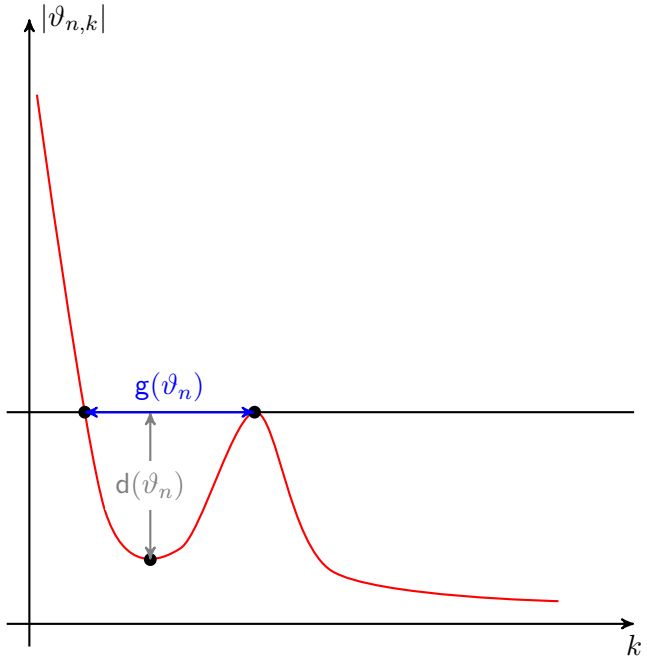

Figure 3: **Idealized scheme of the parametrization of the HF features of a signal in the Fourier Domain.** The parameter $\mathsf{g}(f)$ is the location parameter in the frequency scale which corresponds to the distance of the HF features from the low-frequency components of the signal. The parameter $\mathsf{d}(f)$ is the intensity parameter which corresponds to the relative amplitude of the HF features.

First step: Pre-processing the signal

Replacing x_i^n by $x_i^n + C$ for some arbitrary constant C, with no loss of generality, we may (and will) assume that

$$|\vartheta_{n,0}| > \max_{0\le k\le n-1} |\vartheta_{n,k}|. \tag{3}$$

Condition (3) is in force from now on. We transform $\vartheta_n = (\vartheta_{n,k})_{0\le i\le n-1}$ into a non-decreasing sequence $\mu_n^{(m)} = (\mu_{n,j}^{(m)})_{m\le j\le n-m}$ that depends on a certain smoothing parameter m (with $0 \le m \le n-1$) defined as follows:

$$\mu_{n,m}^{(m)} = \min_k \vartheta_{n,k}^{(m)} \le \mu_{n,m+1}^{(m)} \le \ldots \le \mu_{n,j}^{(m)} \le \mu_{n,n-m}^{(m)} = \max_k \vartheta_{n,k}^{(m)}$$

where

$$\vartheta_{n,k}^{(m)} = \Big(\tfrac{1}{2m+1} \sum_{l=k-m}^{k+m} |\vartheta_{n,l}|^2\Big)^{1/2}, \quad m \le k \le n-m-1. \tag{4}$$

In other words, the sequence $\mu_n^{(m)}$ is the order statistics of a $2m$-regularised version of ϑ_n.

Remark 1. *The smoothing parameter m is needed as soon as the signal observed displays singularities e.g. a jump discontinuity or a fast transition of monotonicity of the trend. These phenomena are approximated by the harmonic sequence $\{e^{j2\pi k\cdot}, k \in \mathbb{Z}\}$, and when projected in the Fourier domain, the amplitude spectrum displays a serie of spikes (cf Figure 4). These phenomena are related to Gibbs phenomenon ([24], chapter 2) and give rise to spikes in the Fourier domain which can be falsely interpreted as HF features. The regularization with an adequate choice of the parameter m solves this issue (cf Figure 4 and Section 3).*

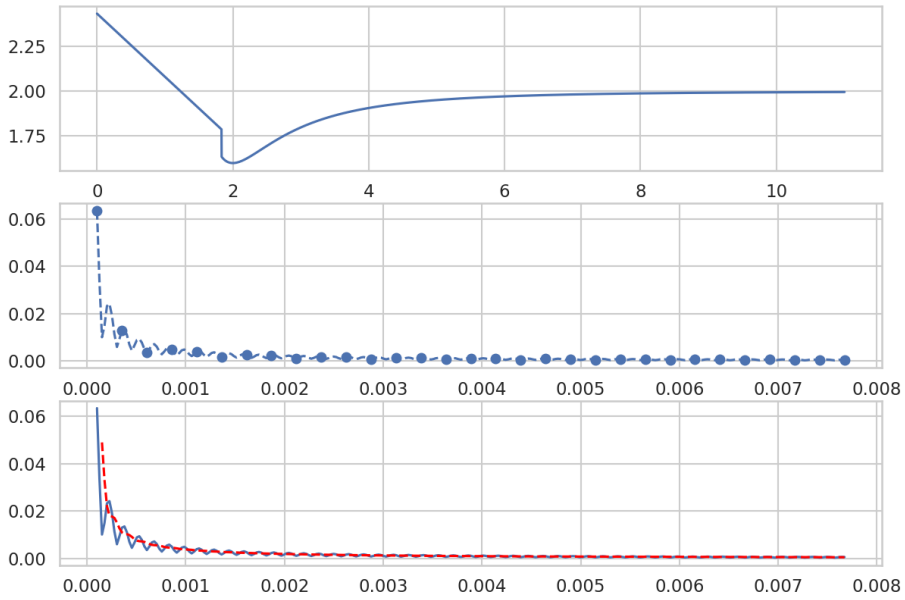

Figure 4: **Graph of a test signal with a jump and a change of monotonicity and its single-sided amplitude spectrum. Top:** Graph of the signal with a decreasing, increasing and stationary part. **Middle:** Zoom on the low frequency of the amplitude spectrum for $n = 10000$ samples of the signal. The blue dot markers emphasize one over ten samples of the signal. **Bottom:** Plot of the amplitude spectrum of the test signal (plain line). The dash line corresponds to the plot of $(\vartheta_{n,k}^{(3)})_{3\leq k\leq n-4}$ defined by (4) with $m = 3$.

Remark 2. *The regularisation of order $2m$ transforms the sequence ϑ_n of n terms into a sequence of $n - 2m$ terms in order to avoid boundary effects. We label the indices of the series*

from m to $n-m-1$ in so that the parameter k in $\vartheta_{n,k}^{(m)}$ is reminiscent of a frequency parameter and we formally have $\vartheta_{n,k}^{(0)} = |\vartheta_{n,k}|$.

Second Step: Detection and Localization of significant features in the Fourier domain.

Define, for $x \geq 0$

$$\mathsf{a}(x) = \mathsf{a}_n^{(m)}(x) = \min\left\{k \mid m \leq k \leq n-m-1,\ \vartheta_{n,k}^{(m)} \leq x\right\} \tag{5}$$

and

$$\mathsf{b}(x) = \mathsf{b}_n^{(m)}(x) = \max\left\{ \arg\max\left\{\vartheta_{n,k}^{(m)} \mid \mathsf{a}(x) \leq k \leq n-m-1\right\}\right\}. \tag{6}$$

Remark 3. *The index $\mathsf{a}(x)$ is the minimal frequency at which searching for HF features starts, getting rid of the potentially high energy levels arising from the low frequency part of the signal. The index $\mathsf{b}(x)$ is a maximal frequency for which the energy level x is reached in the search zone $\{\mathsf{a}(x), \mathsf{a}(x)+1, \ldots, n-m\}$.*

Define the sets

$$\mathcal{A}_n^{(m)} = \left\{\mu_{n,j}^{(m)} \mid \mu_{n,j}^{(m)} = \vartheta^{(m)}_{n,\mathsf{b}(\mu_{n,j}^{(m)})},\ m \leq j \leq n-m-1\right\}$$

and

$$\mathcal{S}_n^{(m)} = \left\{\mu_{n,j}^{(m)} \in \mathcal{A}_n^{(m)} \mid \mathsf{b}(\mu_{n,j}^{(m)}) > \mathsf{a}(\mu_{n,j}^{(m)}),\ m \leq j \leq n-m-1\right\}.$$

Remark 4. *The set $\mathcal{A}_n^{(m)}$ represents potential candidates for maximum energy levels of a HF feature, while $\mathcal{S}_n^{(m)}$ represents the set of intensities of the spikes of ϑ_n.*

Third Step: Definition of the HF features parameters.

To define the HF features, we now select in the set $\mathcal{S}_n^{(m)}$ the feature with maximum relative amplitude. Let us define

$$\mathsf{d}(x) = \mathsf{d}_n^{(m)}(x) = x - \min\left\{\vartheta_{n,k}^{(m)} \mid m \leq k \leq \mathsf{b}_n^{(m)}(x)\right\} \tag{7}$$

and we obtain a maximum intensity of HF feature as

$$\iota_n^{(m)}(\vartheta_{n,\cdot}) \in \max\left\{ \arg\max_{x\in\mathcal{S}}\left(x - \min_{m\leq k\leq \mathsf{b}_n^{(m)}(x)} \vartheta_{n,k}^{(m)}\right)\right\} = \max\left\{\arg\max_{x\in\mathcal{S}} \mathsf{d}_n^{(m)}(x)\right\}$$

if $\mathcal{S}_n^{(m)}$ is non empty and $\iota_n^{(m)}(\vartheta_{n,\cdot}) = 0$ otherwise. Moreover if the set $\arg\max_{x\in\mathcal{S}} \mathsf{d}_n^{(m)}(x)$ is not reduced to a singleton taking its maximum ensures us to obtain a unique element for $\iota_n^{(m)}(\vartheta_{n,\cdot})$ i.e. the feature of maximum relative amplitude and maximum intensity. We are ready to give a quantitative definition of a HF feature:

Definition 1. *To any discrete signal* $\vartheta_n = (\vartheta_{n,k})_{0\leq i\leq n-1}$ *given in the Fourier domain, we associate a high-frequency feature (HF feature)* $\big(\mathsf{G}_{n,m}(\vartheta_n), \mathsf{D}_{n,m}(\vartheta_n)\big)$ *at discretisation level* $n \geq 1$ *and smoothing level* $m \leq \frac{n-1}{2}$ *as follows:*

$$\mathsf{G}_{n,m}(\vartheta_n) = \mathsf{b}_n^{(m)}\big(\iota_n^{(m)}(\vartheta_n)\big) - \mathsf{a}_n^{(m)}\big(\iota_n^{(m)}(\vartheta_n)\big)$$

and

$$\mathsf{D}_{n,m}(\vartheta_n) = \mathsf{d}_n^{(m)}\big(\iota_n^{(m)}(\vartheta_n)\big),$$

where $\mathsf{b}_n^{(m)}$, $\mathsf{a}_n^{(m)}$ *and* $\mathsf{d}_n^{(m)}$ *are defined in* (6), (5) *and* (7) *respectively.*

Remark 5. *The parameters* $\mathsf{G}_{n,m}(\vartheta_n)$ *and* $\mathsf{D}_{n,m}(\vartheta_n)$ *are two distances (*$\mathsf{G}_{n,m}(\vartheta_n)$ *is a distance on the frequency axis and* $\mathsf{D}_{n,m}(\vartheta)$ *on the intensity axis). This couple of parameters provides a characterization in the discrete Fourier domain of events defined as HF features. For each signal, the parametric characterization is unique. It describes the peak with the highest distance between its amplitude and the minimum amplitude of the Fourier coefficients of lower frequencies (with* $\mathsf{D}_{n,m}(\vartheta_n)$*). The parameter* $\mathsf{G}_{n,m}(\vartheta_n)$ *gives the distance in frequency indices between the peak and the components in the low frequencies with the same intensity (see Figure 3).*

2 Testing for HF features

We keep-up with the statistical setting introduced in Equation (1): we observe

$$y_i^n = x_i^n + \sigma\xi_i^n, \quad i = 0, \ldots, n-1, \tag{8}$$

where $(x_i^n)_{0\leq i\leq n-1}$ is the signal of interest and the $\sigma\xi_i^n$ are independent centred Gaussian random variables with noise variance σ^2, for some (large) $n \geq 1$, interpreted as a maximal discretisation resolution level or equivalently a maximal frequency of observation. Applying the discrete Fourier transform DFT_n on both sides of (8), we equivalently observe

$$\widehat{\vartheta}_{n,k} = \vartheta_{n,k} + \sigma\widetilde{\xi}_{k,n}, \quad k = 0, \ldots, n-1,$$

where the $\sigma\widetilde{\xi}_{k,n}$ are independent centred Gaussian random variables with variance σ^2 as well, thanks to the fact that DFT_n is an orthogonal linear mapping. From data $(y_i^n)_{0\leq i\leq n-1}$ or rather $(\widehat{\vartheta}_{n,k})_{0\leq k\leq n-1}$, we wish to construct a statistically significant test of the absence of HF feature as the null, against a set of local alternatives where some HF features are present.

2.1 Construction of a statistical test

Thanks to the characterisation of HF features via $\big(\mathsf{D}_{n,m}(\vartheta_n), \mathsf{G}_{n,m}(\vartheta_n)\big)$ given in Definition 1, we test the null

$$\mathcal{H}^0_{n,m,\nu,c}: \quad \mathsf{G}_{n,m}(\vartheta_n) < \nu, \quad \mathsf{D}_{n,m}(\vartheta_n) < c$$

against the local alternatives

$$\mathcal{H}^1_{n,m,\nu,c}:\ \mathsf{G}_{n,m}(\vartheta_n)\geq\nu\ \text{ and }\ \mathsf{D}_{n,m}(\vartheta_n)\geq c$$

where $\nu>0,\ c>0$ are thresholds to determine significant HF features. The null hypothesis $\mathcal{H}^0$, is that there is no significant HF feature in the signal tested. On the contrary, the hypothesis $\mathcal{H}^1$ implies that the signal has significant HF feature. For the test to be powerful, the main problem is to define the couple (ν,c): for too small values any signal shall reject $\mathcal{H}^0$ whereas for large values, any signal shall accept $\mathcal{H}^0$. We obtain simple test statistics for $\big(\mathsf{G}_{n,m}(\vartheta_n),\mathsf{D}_{n,m}(\vartheta_n)\big)$ by setting

$$\widehat{\mathsf{G}}_{n,m}=\mathsf{G}_{n,m}(\widehat{\vartheta}_n)=\mathsf{b}_n^{(m)}\big(\iota_n^{(m)}(\widehat{\vartheta}_n)\big)-\mathsf{a}_n^{(m)}\big(\iota_n^{(m)}(\widehat{\vartheta}_n)\big)$$

and

$$\widehat{\mathsf{D}}_{n,m}=\mathsf{D}_{n,m}(\widehat{\vartheta}_n)=\mathsf{d}_n^{(m)}\big(\iota_n^{(m)}(\widehat{\vartheta}_n)\big).$$

In order to compute the p-value of the test, we design a Monte Carlo procedure simulating a proxy of the data $(y_i)_{0\leq i\leq n-1}$ under the null $\mathcal{H}^0$. Using the proxy, we define a reject region of our test for a risk level α and the p-value of the data $(y_i)_{0\leq i\leq n-1}$.

Rejection zone at risk level α.

We first simulate N times $y^{(0)}_{\lambda,n}$ defined in (14) below, which is a simulated proxy of the data $(y^n_i)_{0\leq i\leq n-1}$ with HF features removed from the signal $(x^n_i)_{0\leq i\leq n-1}$. Repeating independently N times the procedure, we obtain a Monte Carlo sequence

$$y^{(0),k}_{\lambda,n}\ k=1,\dots,N.$$

In a second step, we denote by E^0_N the cloud of points representing the HF features parameters of these simulated signals (with HF features removed but with Gaussian noise):

$$E^0_N=\left\{\left(\mathsf{G}_{n,m}\left(\mathsf{DFT}[y^{(0),k}_{\lambda,n}]\right),\mathsf{D}_{n,m}\left(\mathsf{DFT}[y^{(0),k}_{\lambda,n}]\right)\right)\ |\ k=1,\dots,N\right\}. \tag{9}$$

We define the function $P:\mathbb{R}^2_+\to F\subset[0;1]$:

$$P(g,d)=N^{-1}\sum_{k=1}^{N}\mathbf{1}_{\left\{\mathsf{G}_{n,m}\left(\mathsf{DFT}[y^{(0),k}_{\lambda,n}]\right)\geq g,\ \mathsf{D}_{n,m}\left(\mathsf{DFT}[y^{(0),k}_{\lambda,n}]\right)\geq d\right\}}. \tag{10}$$

Hence $P(g,d)$ is the proportion of points in E^0_N located in the North-East quarter of the plane centered on (g,d) (cf Figure 5). In order to reduce the computation cost, we only consider the restriction of P to the set E^0_N. Thus if E^0_N is reduced to a singleton, then the image set $P(E^0_N)$ is equal to $\{1\}$, on the contrary if E^0_N contains N disjoint points then the minimal bound on $P(E^0_N)$ is $\frac{1}{N}$. For a risk level $\alpha\in P(E^0_N)$, the rejection zone of our test is defined as

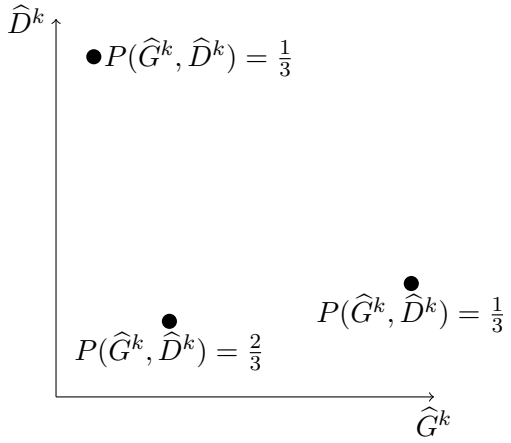

Figure 5: Cloud of points $\left(\widehat{\mathsf{G}}^k, \widehat{\mathsf{D}}^k\right) = \left(\mathsf{G}_{n,m}\left(\mathsf{DFT}[y^{(0),k}_{\lambda,n}]\right), \mathsf{D}_{n,m}\left(\mathsf{DFT}[y^{(0),k}_{\lambda,n}]\right)\right)$ for $k = 1, 2, 3$.

$$\mathcal{R}_{m,n}(\kappa_1^\alpha, \kappa_2^\alpha) = \left\{(y_i)_{1\le i\le n} \text{ defined by (1) s.t. } \widehat{\mathsf{G}}_{n,m} \ge \kappa_1^\alpha, \ \ \widehat{\mathsf{D}}_{n,m} \ge \kappa_2^\alpha\right\} \tag{11}$$

where $(\widehat{\mathsf{G}}_{n,m}, \widehat{\mathsf{D}}_{n,m})$ is the test statistics and $\left(\kappa_1^\alpha, \kappa_2^\alpha\right) \in E_N^0$ are such that

$$P(\kappa_1^\alpha, \kappa_2^\alpha) = \alpha. \tag{12}$$

Remark 6. *The risk level α is imposed by the Monte Carlo sequence, $\alpha \in P(E_N^0) \subset [\frac{1}{N}; 1]$. For example, Figure 5 represents an arbitrary set E_N^0 for $N = 3$. Consequently, we note that $\frac{1}{3} \le \alpha \le 1$ in order to obtain candidates $\kappa_1^\alpha, \kappa_2^\alpha$. For $\alpha < \frac{1}{3}$ no candidate can be obtained by this procedure and its associated reject region is not defined. Moreover there can be multiple reject regions defined for the same risk level α (in the example whe have two reject regions for $\alpha = \frac{1}{3}$).*

The main idea behind the computation of the couples $\left(\mathsf{G}_{n,m}\left(\mathsf{DFT}[y^{(0),k}_{\lambda,n}]\right), \mathsf{D}_{n,m}\left(\mathsf{DFT}[y^{(0),k}_{\lambda,n}]\right)\right)$ is to generate random outcomes under the null $\mathcal{H}^{(0)}$ that enable us to compute risk level by Monte Carlo. The couples correspond to the relative amplitude and the frequency gap for a non-oscillating signal with noise. We also get reject region(s) of level α thanks to the threshold(s) $\left(\kappa_1^\alpha, \kappa_2^\alpha\right)$. We do not need uniqueness of the reject region in order to define and compute the p-value, see below.

Definition of the $\mathrm{p}-$value.

The $\mathrm{p}-$value of the observations $(y_i^n)_{0\le i\le n-1}$ is defined as

$$\mathrm{p}-\mathrm{value}\left((y_i^n)_{0\le i\le n-1}\right) = \min\left\{\alpha \in P(E_N^0) \mid \widehat{\mathsf{G}}_{n,m} \ge \kappa_1^\alpha, \quad \widehat{\mathsf{D}}_{n,m} \ge \kappa_2^\alpha\right\}. \tag{13}$$

An equivalent definition of the p-value of the observations $(y_i^n)_{0\le i\le n-1}$ is obtained via

$$\mathrm{p}-\mathrm{value}\big((y_i^n)_{0\le i\le n-1}\big) = \inf\big\{\alpha\in P(E_N^0)|\ (y_i^n)_{0\le i\le n-1}\in\mathcal{R}_{m,n}\big(\kappa_1^\alpha,\kappa_2^\alpha\big)\big\}.$$

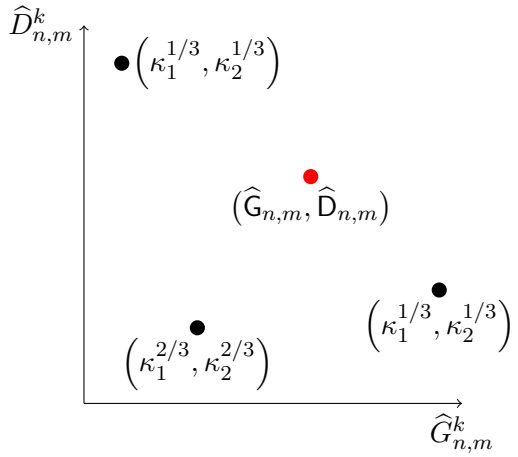

Figure 6: **Point cloud $\left(\widehat{\mathsf{G}}^k_{n,m},\widehat{\mathsf{D}}^k_{n,m}\right)$ for $k=1,2,3$ (black dots) and HF feature parameters $\left(\widehat{\mathsf{G}}_{n,m},\widehat{\mathsf{D}}_{n,m}\right)$ (red dot).**

Remark 7. *The computation of the* p$-$*value is illustrated schematically in Figure 6. We observe the point cloud formed by $\left(\widehat{\mathsf{G}}^k_{n,m},\widehat{\mathsf{D}}^k_{n,m}\right)$ (the black dots) for $k=1,2,3$. On the vertical axis we have the relative amplitude and on the horizontal axis we have the gap in frequency between the oscillations and the trend in the Fourier domain. The red dot illustrates the HF features parameters subject to the test. We use the $\left(\widehat{\mathsf{G}}^k_{n,m},\widehat{\mathsf{D}}^k_{n,m}\right)$ as our grid to compute the p-value. Using (12), we compute the level α and obtain consequently the $(\kappa_1^\alpha,\kappa_2^\alpha)$ for each element of the grid. The p-value is $\frac{2}{3}$ in this example.*

The p $-$ value gives a confidence index for non-rejecting the null. This index is meaningful provided the test has a good power, *i.e.* if the probability of making a type II error is small. Hence the p $-$ value of $\big((y_i^n)_{0\le i\le n-1}\big)$ is our measure of confidence in non-rejection of the null $\mathcal{H}^0$. The main difficulty however lies in solving (11) since ϑ_n remains unknown under the null and that there are no reason that $\widehat{\mathsf{G}}_{n,m}$ or $\widehat{\mathsf{D}}_{n,m}$ are pivotal statistics under the null. We describe below a numerical procedure based on Monte Carlo simulation that estimates $y^{(0)}_{\lambda,n}$ a proxy of the data with HF features removed but with noise.

2.2 A Monte Carlo procedure for the simulation of the null

In order to evaluate (11) and (13), we first build a low-frequency estimator $\widehat{x}^{(0)}_{\lambda,n}$ from the data $(y^n_i)_{0\le i\le n-1}$ that removes the potential HF features. The estimator depends on a regularisation parameter λ. We next define

$$y^{(0)}_{\lambda,i,n} = \widehat{x}^{(0)}_{\lambda,i,n} + \widehat{\sigma}_n \epsilon^n_i, \quad i = 0, \dots, n-1, \tag{14}$$

where the ϵ^n_i are independent centred Gaussian random variables that we simulate and $\widehat{\sigma}_n$ is an estimator of the standard deviation of the noise. The simulated signal $(y^{(0)}_{\lambda,i,n})_{0\le i\le n-1}$ obtained by estimating a proxy of f with HF features removed with additional simulated noise serves as a proxy of the data $(y^n_i)_{0\le i\le n-1}$ under the null $\mathcal{H}^0$.

Numerical computation of $\widehat{f}^{(0)}_{\lambda,n}$

Trend estimation or filtering for mimicking a signal with HF features removed has many applications and hence it has been extensively studied. It has given rise to the smoothing and filtering methods such as the moving average [23], smoothing splines [21], Hodrick-Prescott filtering [20], ℓ_1-trend filtering [12] and so on. The trend is considered as the general shape of a signal or a time series. Although the trend is often understood and perceived intuitively, its estimator relies on the definitions given to the trend. The differences between the various definitions of the trend are a matter of interpretation. Considering the different definitions of the trend, the choice of the method to estimate this component is more likely qualitative. In the following, the trend is considered as the underlying slowly varying component of the signal and we choose the ℓ_1-trend filtering method described in [12] to estimate it. The estimator of $\widehat{x}^{(0)}_{\lambda,n}$ as a n-dimensional vector is then the solution of the following optimisation problem:

$$\widehat{x}^{(0)}_{\lambda,n} \in \operatorname*{arg\,min}_{x\in\mathbb{R}^n} \frac{1}{2}\sum_{i=1}^{n-1}(y^n_i - x^n_i)^2 + \lambda \sum_{i=1}^{n-2} |x^n_{i-1} - 2x^n_i + x^n_{i+1}|, \tag{15}$$

where $\lambda \ge 0$ is a regularisation parameter which controls the trade-off between the smoothness of $\widehat{x}^{(0)}_{\lambda,n}$ and the residual $\sum_{i=0}^{n-1}\left(\widehat{x}^{(0)}_{\lambda,n,i} - y^n_i\right)^2$. We note that the second term $\sum\limits_{i=1}^{n-2} |x^n_{i-1} - 2x^n_i + x^n_{i+1}|$ is the ℓ^1-norm of the second order variations of the sequence (x^n) (i.e. the discretization of the corresponding L^1-norm of the second derivative of a function). Moreover, for any sequence (x^n),

$$|x^n_{i-1} - 2x^n_i + x^n_{i+1}| = 0,\ \forall i = 0, \dots, n-1 \iff x^n_i = \alpha i + \beta, \text{ with } \alpha, \beta \in \mathbb{R},\ \forall i = 0, \dots, n-1.$$

Thus only an affine function has its ℓ^1-norm equal to 0. Hence this method gives an estimator of the trend such that:

(i) $\widehat{x}^{(0)}_{\lambda,n}$ is computed numerically in $\mathcal{O}(n)$ operations,

(ii) as $\lambda \to 0$, $\max_{0\leq i\leq n-1} |\widehat{x}^{(0)}_{\lambda,n} - y^n_i| \to 0$, the estimator converges to the original data,

(iii) as $\lambda \to \infty$, the estimator converges to the best affine fit of the observations. This convergence happens for a finite value of λ [12].

(iv) $\widehat{x}^{(0)}_{\lambda,n}$ is piecewise linear, i.e. there are indices $0 = j_1 < j_2 < \ldots < j_K = n-1$ for which:

$$\widehat{x}^{(0)}_{\lambda,n,i} = \alpha_k i + \beta_k, \quad j_k < i < j_{k+1}, \quad k = 1, \ldots, K-1.$$

The ℓ_1-trend filtering method is well suited to extract the trend components of the signals studied in Section 3. Since the signals display singularities such as discontinuous jumps, the trend extracted is well approximated by a piecewise linear function. Moreover the HF features in the signals are components looking like sine waves and varying at an intermediate pace. However interpolating a sine wave by a piecewise linear function requires a fine scale and thus the parameter λ has to be close to 0. Rising slightly the value of λ allows us to capture the trend without the HF features. Moreover there exists a threshold $\lambda_{\max} \in \mathbb{R}_+$ [12] such that $\widehat{x}^{(0)}_{\lambda_{\max},n}$ is the trend estimator corresponding to the best affine fit. It implies that the choice of λ is restricted to the bounded open interval $(0,\ \lambda_{\max})$. Since there is no optimal criterium to choose λ, the choice of the parameter is qualitative and motivated empirically (see Section 3).

Numerical estimation of the noise level $\widehat{\sigma}_n$

The estimator of the standard deviation of the noise is the second ingredient needed in order to compute $\widehat{f}^{(0)}_{\lambda,n}$ in (14). The methods to estimate the level of noise are closely linked to the methods of signal denoising and thus have been extensively studied. The method chosen to estimate the noise level is the median absolute deviation and the denoised signal is obtained thanks to the wavelet shrinkage methods [5, 6, 7, 8].

We assume that our data $y = (y_i)_{0\leq i\leq n-1}$ are such that $n = 2^{J+1}$ for $J > 0$. We then consider an orthogonal wavelet transform matrix $\mathcal{W}$ for a given filter. Choosing wavelets (e.g. Coiflet, Daubechies, Haar) and varying the combinations of parameters M (number of vanishing moments), S (support width) and j_0 (low-resolution cut-off) one may construct various orthogonal matrices $\mathcal{W}$ (see for details [15], chapter 7). In this paper we use the Symmlet with parameter 8 which has $M = 7$ vanishing moments and support length $S = 15$. The wavelet coefficients of y are denoted by w and

$$w = \mathcal{W}x + \sigma\widetilde{\xi},$$

where $\widetilde{\xi} = \mathcal{W}\xi$ is a standard Gaussian random vector by orthogonality of $\mathcal{W}$. For convenience, we index dyadically the vector of the wavelet coefficients

$$w_{j,k} \quad j = 0, \ldots, J, \quad k = 0, \ldots, 2^j - 1.$$

We make the legitimate assumption that empirical wavelet coefficients at the finest resolution level J are essentially pure noise. Hence the standard deviation estimator $\widehat{\sigma}_n$ is the median absolute deviation

$$\widehat{\sigma}_n = \frac{\text{median}(w_{J,\cdot})}{\Phi^{-1}(3/4)}, \tag{16}$$

where $\Phi^{-1}(\cdot)$ is the inverse of the cumulative distribution function for the standard normal distribution. Thus $\widehat{\sigma}_n$ is a consistent estimator of σ. It is interesting to note that further computations give the VisuShrink estimator $\widehat{x}_n$ of the signal $(x_i^n)_{0\leq i\leq n-1}$

$$\widehat{x}_n = \mathcal{W}^T \bullet \widehat{w}^{n,j_0} \bullet \mathcal{W}, \tag{17}$$

where j_0 denotes a low resolution cut-off and $\widehat{w}^{n,j_0}$ is the estimator in the wavelet domain

$$\widehat{w}^{n,j_0} = \begin{cases} w_{j,\cdot} & j < j_0 \\ \text{sign}(w_{j,\cdot})\left(|w_{j,\cdot}| - \widehat{\sigma}_n(2\log n)^{1/2}\right)_+ & j_0 \leq j \leq J \end{cases}.$$

The first reason that motivated this choice is that the shrinkage methods attempt to remove whatever noise is present and retain whatever signal is present regardless of the frequency [8]. The goal of this paper is to estimate HF features in noisy signals. However the traditional methods of noise removal such as low-pass filters are based on frequency-dependent estimators, which can also impact and distort the results of the HF feature procedure. The second reason is that these methods are data-driven and no specific assumptions on the signal are required. The wavelet shrinkage is spatially adapted and the method is efficient for a wide variety of signals even when the signals exhibit spatial inhomogeneities [8]. Finally these methods are proven to be nearly optimal for the mean squared error criterion when the smoothness of the original signal is unknown [7].

3 Simulation example: sanity check of the procedure.

Pre-processing: a data-driven choice of m

We first address the delicate issue of choosing the smoothing parameter m. Define a sequence $(m_i)_{1\leq i\leq K}$ such that

$$1 = m_1 < m_2 < \ldots < m_K \leq \tfrac{n-1}{2}.$$

We can take for instance $m_i = i$ for $i = 1, \ldots, K$. Note that $K \in \{1, \ldots, \frac{n-1}{2}\}$ is the parameter defining the length of the finite sequence $(m_i)_{1\leq i\leq K}$. This parameter can be fixed by the user in order to reduce the number of iterations of the procedure to compute the HF features. However, a standard choice of K to obtain a data-driven procedure is $K = \frac{n-1}{2}$, since averaging the signal over more than half of the sample size is obviously meaningless. A good rule of thumbs is that

$K = n^{\frac{1}{2}}$, since it reduces the number of calculations and remains pertinent compared to the range of the signal. Let

$$i^\star \in \underset{1\leq i\leq K}{\arg\max} \left|\widehat{\mathsf{G}}_{n,m_i} - \widehat{\mathsf{G}}_{n,m_{i-1}}\right|,$$

then

$$\widehat{m} = \begin{cases} m_{i^\star} & \text{if} \quad \widehat{\mathsf{G}}_{n,m_{i^\star}} > \widehat{\mathsf{G}}_{n,m_{i^\star-1}} \\ m_{i^\star-1} & \text{otherwise.} \end{cases} \tag{18}$$

As previously stated in Remark 1, the empirical signals observed are non-monotonous, contain singularities and transient oscillations. Their amplitude spectra display a series of spikes in the low-frequencies and in the mid or high frequencies. Hence without a pre-processing step, the HF feature parameters (Definition (1)) characterize the low frequencies features (i.e. the trend represented in the amplitude spectrum by spikes in the low frequencies, see Figure 4).
In order to solve this problem, we regularize the Fourier coefficients as defined in (4). The sequence $(m_k)_{1\leq k\leq K}$ gradually smoothes the Fourier amplitude spectrum: the spikes in the low frequencies merge together whereas the isolated spikes in the mid or high frequencies (corresponding to transient oscillations) slightly decrease in amplitude but remain significant. The data-driven choice of m is well adapted to regularize the empirical signals since it chooses the parameter $\widehat{m}$ from the sequence $(m_k)_{1\leq k\leq K}$ which maximizes the difference between the localisation parameters $\widehat{\mathsf{G}}$ for two consecutive smoothing parameters. Thus the spikes located in a close frequency range have been smoothed and the remaining spikes of significant amplitude for the regularization parameter $\widehat{m}$ are isolated in the Fourier amplitude spectrum.

Defining a test signal

To study numerically the validity of the procedure and the statistical test, we first compute a simulated signal where all the parameters are known. To do so, we superimpose three signals: one for the general trend of the curve, one for the HF features, and one for the noise. The signal obtained is the vector $(S_i)_{0\leq i\leq n-1}$:

$$S_i = T_i + O_i + \sigma\xi_i, \tag{19}$$

where $\sigma > 0$ is the parameter corresponding to the level of noise and ξ_i are realizations of independent and identically normally distributed random variables. Moreover $(T_i)_{0\leq i\leq n-1}$ corresponds to the trend and $(O_i)_{0\leq i\leq n-1}$ to the HF features (cf Figure 7).

For the general trend, we choose the Lennard Jones potential [11], since we notice that its DFT is not monotonously decreasing in the low frequency range (see Figure 2) and that it displays a similar shape as the experimental signals presented in Section 4. The Lennard Jones potential is defined by P_i:

$$P_i = \left(c_1\left[\left(\frac{c_2}{i}\right)^p - c_3\left(\frac{c_2}{i}\right)^q\right] + c_4\right).$$

Since this potential is not defined at 0, we link the potential to an affine function. Hence we introduce the index j $(0 < j < n-1)$ which connects the potential to the affine function. We denote the trend by the vector $(T_i)_{0\leq i\leq n-1}$:

$$T_i = \left(\frac{P_{j+1} - P_j}{j+1} i + P_j\right) \mathbb{1}_{\{0\leq i\leq j\}} + P_i \mathbb{1}_{\{j+1\leq i\leq n-1\}}.$$

The HF features in the test signal correspond to sine waves and are located at a specific time

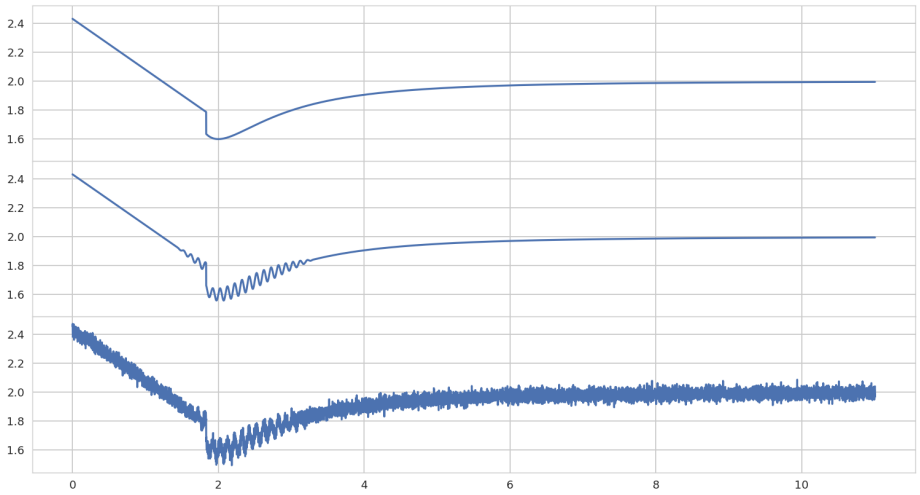

Figure 7: **Simulation of the test signal defined by** (19)**.**The x-axis is the time in hours. (Up) Plot of $(T_i)_{0\leq i\leq 10^5}$ with parameters $c_1 = 0.4$, $c_2 = c_3 = c_4 = 2$, $\frac{p}{2} = q = 3, j_0 = 1700$, $j_1 = 3400$. (Middle) Plot of $(T_i + O_i)_{0\leq i\leq 10^5}$ with the same parameters and $c_a = 0.05$, $c_f = 10$. (Down) Plot of $(S_i)_{0\leq i\leq 10^5}$ with the same parameters and $\sigma = 0.025$.

interval. Hence we introduce the indices $0 < j_0 < j_1 < n-1$ which localize the oscillations in the signal, and we define the oscillations by the vector $(O_i)_{0\leq i\leq n-1}$:

$$O_i = c_a(i - j_0)(j_1 - i)\sin(2\pi c_f i)\left(\frac{4}{(j_1 - j_0)^2}\right)\mathbb{1}_{\{j_0\leq i\leq j_1\}} \tag{20}$$

where c_a (resp. c_f) is the parameter for the amplitude (resp. the frequency) of the oscillations.

σ	$\frac{1}{10}c_a$	$\frac{1}{2}c_a$	c_a	$2c_a$	$10c_a$
$\widehat{m}$	6	6	6	6	24
$\widehat{\mathsf{G}}_{n,\widehat{m}}$ (Hz)	2.069e−3	2.044e−3	2.145e−3	1.943e−3	8.437e−2
$\widehat{\mathsf{D}}_{n,\widehat{m}}$	1.73e−4	1.807e−4	1.807e−4	1.844e−4	2.768e−3
p-value	5e−5	5e−5	5e−5	5e−5	4.023e−1

Figure 8: **Table of estimators and p-values of the sanity-check signals.** The simulation of the null is performed with the real trend of the signals.

Numerical computations and robustness of the procedure. We want to understand the robustness of the numerical procedure when the frequencies and the amplitudes of the oscillations are fixed but the level of noise varies. Other said, for which parameters of the oscillations and for which level of noise does the test return that the signal oscillates (or not)? In order to answer this question, we propose the following sensitivity analysis.

First we remind the parameters in our system. From the signal construction, we have three parameters :

- σ the standard deviation of the normal distributed noise,
- c_a the parameter corresponding to the amplitude of the oscillations,
- c_f the parameter corresponding to the frequency of the oscillations (since the time scale is in hours, $c_f/3600$ is expressed in Hz).

The smoothing parameter $\widehat{m}$ is chosen thanks to the data-driven procedure described previously (18).

The relevant output of our model is the p-value of the signals computed thanks to the numerical procedure. A natural way to study the sensitivity of the p-value to the parameters is to fix all parameters but one and observe the effect on the p-values obtained. In this example the varying parameter is the level of noise $\sigma \in \left\{\frac{1}{10}c_a,\ \frac{1}{2}c_a,\ c_a,\ 2c_a,\ 10c_a\right\}$.

First sanity check test

Since we are working with a constructed sanity check signal, we obtained $(\widehat{\mathsf{G}}^k_{n,\widehat{m}}, \widehat{\mathsf{D}}^k_{n,\widehat{m}})$ in Figure 9 by applying the procedure of detection of the HF feature parameters setting $c_a = c_f = 0$ (it

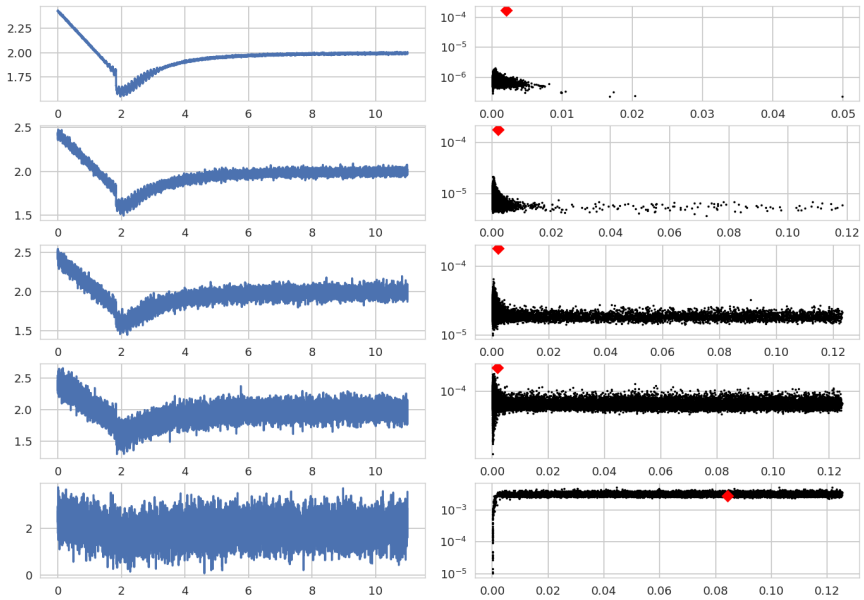

Figure 9: **Numerical results of the procedure on the sanity-check signal when the simulation of the null is performed with the real trend.** (Left column) Plot of $(S_i)_{1\leq i\leq 10^5}$ (19) with the parameters $c_1 = 0.4,\ c_2 = c_3 = c_4 = 2,\ \frac{p}{2} = q = 3, t_0 = 1.43,\ t_1 = 3.30,\ c_a = 0.05,\ c_f = 10$ and $\sigma \in \left\{\frac{1}{10}c_a,\ \frac{1}{2}c_a,\ c_a,\ 2c_a,\ 10c_a\right\}$ from top to bottom. The x-axis is the time in hours. (Right column) The black dots are the cloud of points of the simulation of the null, for $N = 20000$. The red diamond corresponds to the HF features parameters of the corresponding signal on the left column. The x-axis is the localization parameters $\widehat{G}_{n,\widehat{m}}$ and the y-axis is the relative amplitude $\widehat{D}_{n,\widehat{m}}$.

corresponds to $S_i = T_i + \sigma\xi_i$ in (19)). Thus the simulation of the null in Section 2.2 is performed using the real trend of the signal in (14). Then the signal tested (Figure 9) are constructed signal with parameters $c_a = 0.05$, $c_f = 10$ and $\sigma \in \left\{\frac{1}{10}c_a,\ \frac{1}{2}c_a,\ c_a,\ 2c_a,\ 10c_a\right\}$ in (19). The results of the detection of HF features and the statistical test are in Table 8. We note that for standard deviations of the noise between a tenth and the double of the amplitude of the oscillations, the p-value of the test is equal to $5\mathrm{e}-5$. Hence, we are inclined to reject the hypothesis $\mathcal{H}^0$ which corresponds to the event that the signal displays no oscillations. Moreover we note that the signals with standard deviations of the noise between $\frac{1}{10}c_a$ and $2c_a$ have almost the same HF feature parameters where $(\widehat{\mathsf{G}}_{n,\widehat{m}}, \widehat{\mathsf{D}}_{n,\widehat{m}}) \approx (2\mathrm{e}-3, 1.8\mathrm{e}-4)$. In contrast, for the signal with the standard deviation of the noise of $10c_a$, the p-value is equal to 0.4, hence we are inclined to accept that the signal has not significant enough HF feature.

Second sanity check test

The second step is to test the procedure on the same signals but using the trend estimate given by (15) and the noise estimation procedure described in the first step of Section 2.2. The method chosen to estimate the trend of the signal is the ℓ_1-trend filtering [12]. As displayed in Figure 10, the trend estimation is less robust as the standard deviation of the noise rises. However this method is qualitatively the right one to estimate the trend of a signal displaying jumps or spikes.

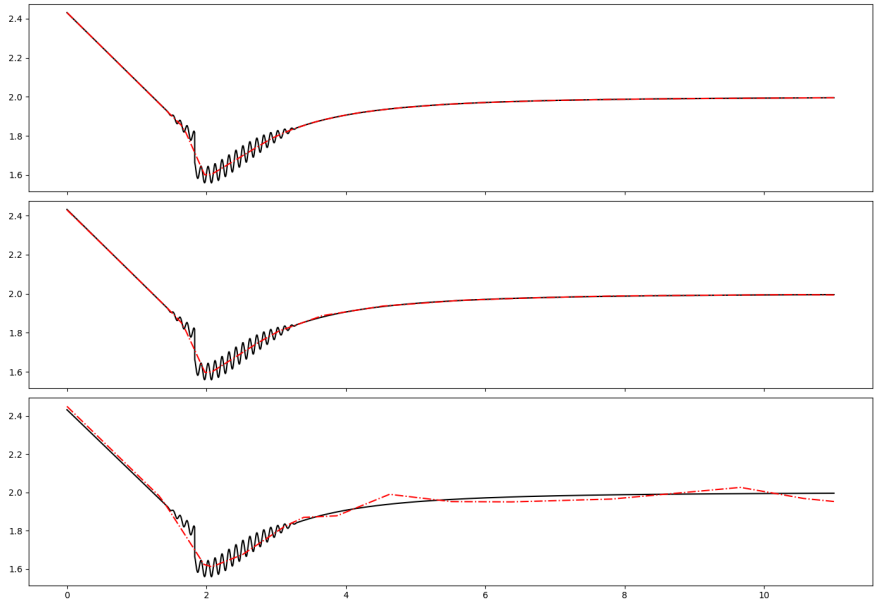

Figure 10: **Numerical estimation of the trend on the sanity check signals.** The x-axis is the time in hours.The parameter in the ℓ_1-trend filtering is $\lambda = 301$. (Up) Plot of $(P_i + O_i)_{0\leq i\leq 10^5}$ in (19) with parameters $c_1 = 0.4,\ c_2 = c_3 = c_4 = 2,\ \frac{p}{2} = q = 3, t_0 = 1.43,\ t_1 = 3.30\ c_a = 0.05,\ c_f = 10$. The dashed line is the ℓ_1-trend estimator when $\sigma = \frac{1}{10}c_a$. (Middle) The dashed line is the ℓ_1-trend estimator when $\sigma = c_a$. (Down)The dashed line is the ℓ_1-trend estimator when $\sigma = 10c_a$.

Hence we compute the procedure to obtain the HF features parameters for the sanity check signals using (19) with standard deviation level $\sigma \in \left\{\frac{1}{10}c_a,\ \frac{1}{2}c_a,\ c_a,\ 2c_a,\ 10c_a\right\}$. The p-values are computed using the ℓ_1-trend estimators in order to obtain the couples $(\widehat{\mathsf{G}}^k_{n,\widehat{m}}, \widehat{\mathsf{D}}^k_{n,\widehat{m}})$ where $k = 1, \ldots, 20000$. The results are in Table 11. Similarly to the first sanity check, the p-values for the signals with a level of noise from $\frac{1}{10}c_a$ to $2c_a$ is equal to $5\mathrm{e}-5$. Hence the procedure detect significant HF features where $\widehat{\mathsf{G}}_{n,\widehat{m}} \approx (2\mathrm{e}-3, 2\mathrm{e}-4)$. Also for a standard deviation of the noise of $10c_a$, the p-value is $5.32\mathrm{e}{-2}$, so that HF feature parameters are not significant enough.

σ	$\frac{1}{10}c_a$	$\frac{1}{2}c_a$	c_a	$2c_a$	$10c_a$
$\widehat{m}$	3	3	3	3	18
$\widehat{\mathsf{G}}_{n,\widehat{m}}$ (Hz)	2.095e−3	2.095e−3	2.044e−3	2.12e−3	1.181e−1
$\widehat{\mathsf{D}}_{n,\widehat{m}}$	1.768e−4	1.784e−4	1.918e−4	2.394e−4	3.593e−3
p-value	5e − 5	5e−5	5e−5	5e−5	5.32e−2

Figure 11: **Table of estimators and p-values of the sanity-check signals.** The simulation of the null is performed with the ℓ_1-estimate of the trend (15) of the signals.

4 Empirical analysis on biological data

The Prion diseases, also known as transmissible spongiform encephalopathies (TSEs), are a group of animal and human brain diseases. The neurodegenerative processes are poorly understood and hence fatal. However the largely accepted hypothesis suggests that the infectious agent (PrPsc) is the misfolded form of the normal Prion protein (PrPc). The PrPsc forms multimeric assemblies (fibrils) which are the prerequisite for the replication and propagation of the diseases [19]. To follow the aggregation kinetics of these fibrils, compare it to mathematical models and get a better understanding of these diseases, several experimental and measurement devices are used, among which the Static Light Scattering (SLS). The Static Light Scattering (SLS) signal is an experimental measurement which describes the temporal dynamics of PrP amyloid assemblies formed in vitro [13] see Fig. 1 taken from [9] (see Appendix A).These signals correspond to an affine transformation of the second moment of the size distribution of protein polymers or fibrils through time [18]:

$$\sum_{i\in\mathcal{I}} i^2 c_i(t) + \sigma,$$

where $\mathcal{I}$ denotes the set of the sizes of the fibrils, c_i the concentration of fibrils of size i which is varying with the time t and $\sigma > 0$ is the experimental noise (σ can be time-dependent). At the beginning of the experiment the fibrils are large, containing in average several hundreds of monomers, which undergo an overall depolymerization process and leads to a decay in the signal. The experiment is carried out with six initial concentrations of fibrils (Figure 12) ranging from 0.25μmol to 3μmol; at higher initial concentrations (0.5μmol and higher), a re-polymerisation process can be observed, which may be viewed by the fact that the trend of the signal increases again before reaching a plateau. Moreover the SLS signals differ in terms of variance of noise and amplitude of oscillations (noticed by sight). We thus study each signal independently.

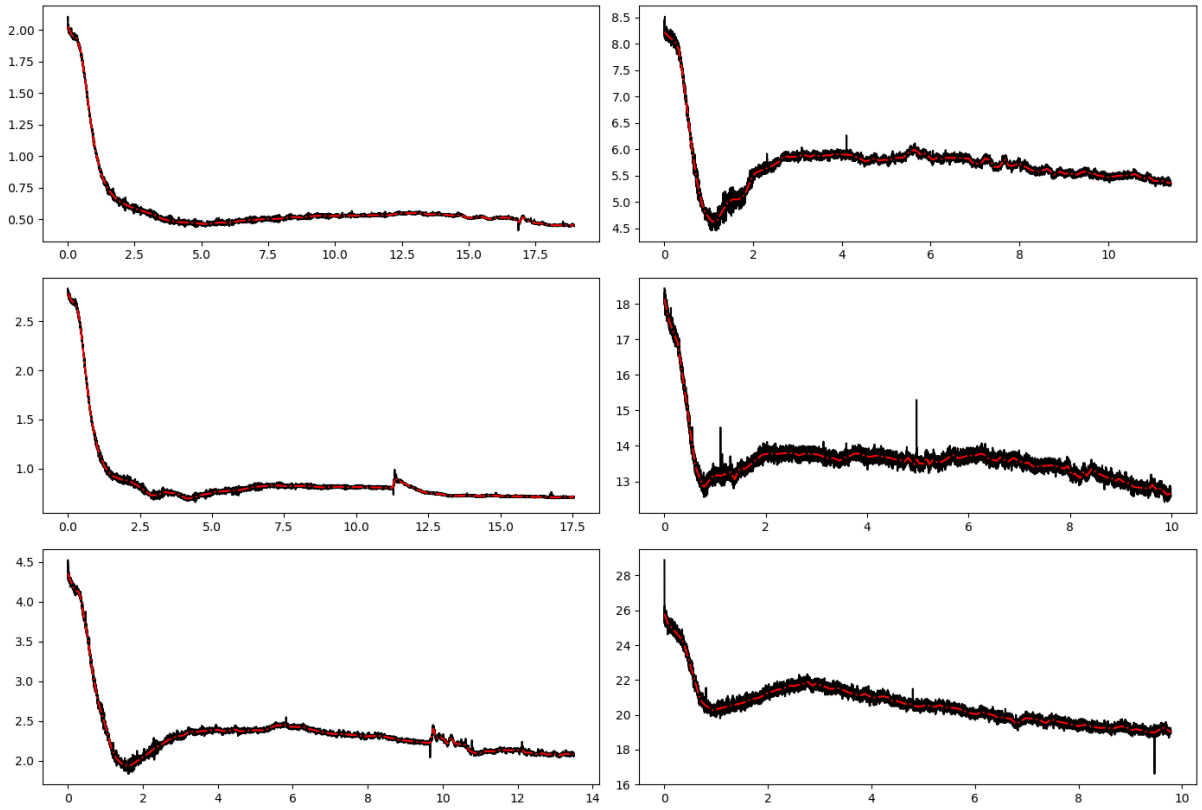

Figure 12: **SLS experiments and trend estimates.** The x-axis is the time in hours. The parameter in the ℓ_1-trend filtering is $\lambda = 31$. (Top left) Plot of $n = 32768$ samples of SLS outputs with initial concentration (I_0) of 0.25μmol of PrP^{Sc} fibrils.The dashed line is the ℓ_1-trend estimator. (Middle left) $I_0 = 0.35\mu$mol (Bottom left) $I_0 = 0.5\mu$mol. (Top right) $I_0 = 1\mu$mol. (Middle right) $I_0 = 2\mu$mol. (Bottom right) $I_0 = 3\mu$mol.

In order to test whether the signals display HF features, we submit the observations to the statistical test described above. The denoised signal and hence the standard deviation of the noise are estimated thanks to the VisuShrink method and the median absolute deviation (cf [8], [6]) using the symmlet wavelet with 8 vanishing moments and the library Wavelab [4] (the same results have been obtained with the homemade python library, see Appendix B). The trend of the signal is estimated with the ℓ_1-trend filtering method with the parameter $\lambda = 31$ (λ is fixed qualitatively in order for the trend to include the discontinuous jumps of the SLS experiments). The results of the statistical test are summarized in Table 13.

We note that all signals display oscillations more or less pronounced (cf. Figure 14). The relative amplitude of the oscillations $\widehat{\mathsf{D}}_{n,\widehat{m}}$ differs from one signal to another for three reasons.

Concentration (μmol)	0.25	0.35	0.5	1	2	3
$\widehat{\sigma}$	$3.553e-3$	$4.72e-2$	$1.11e-2$	$3.09e-2$	$8.44e-2$	$1.287e-1$
$\widehat{m}$	4	3	5	7	9	7
$\widehat{\mathsf{G}}_{n,\widehat{m}}$ (Hz)	4.954e−3	7.53e−3	5.656e−3	8.375.e−3	2.698e−3	4.971e−3
$\widehat{\mathsf{D}}_{n,\widehat{m}}$	9.649e−6	1.863e−5	1.012e−4	6.526e−4	3.345e−4	1.01e−3
p-value	$5e-5$	5e−5	5e−5	5e−5	5e−5	5e−5

Figure 13: **Table of estimators and p-values for the test of presence of HF features in the SLS experiments**

First of all, each signal corresponds to an experiment with a specific initial concentration. The calibration of the experiments is not identical for experiments with different initial concentrations. Secondly, the signals are not on the same scale. The signal with initial concentration of 0.25μmol goes from 0.5 to 2.2 in amplitude, and the signal of initial concentration of 3μmol goes from 16 to 28 in amplitude. Finally, they do not have the same regularization coefficient $\widehat{m}$.

However the frequency localization parameters are comparable. In Table 13, we note that the parameters $\widehat{\mathsf{G}}_{n,\widehat{m}}$ are in the same range of value with a factor of less than 4 between the minimum and maximum $\widehat{\mathsf{G}}_{n,\widehat{m}}$. Finally all the p-value of the tests are equal to $5e-5$, the tests confirm that the signals display significant HF features.

Through this study, we demonstrated the existence of oscillatory behavior in the SLS experiments. The immediate biochemical consequences are the coexistence of structurally distinct PrP assemblies within the same media and the unstable behavior, i.e. out of the thermo-dynamical equilibrium, of the chemical system formed by theses assemblies. Indeed the observation of oscillations in these light-scattering experiments has shed light on the existence of a complex chemical reaction network beyond the existing aggregation-fragmentation models. This has paved the way for new mechanistic models, e.g. a system of reactions which possibly involve several conformations of PrP assemblies [9], capable of explaining such phenomena. Also it has been reported that the existence of multiple conformations of PrP assemblies within an isolate contributes to the adaptation and evolution of Prion as a pathogen to a new environment and a new host [14].

Further biochemical characterizations are required to explore the dynamics of these oscillations and to establish more precise kinetic models. The methodology developed in the present work will lead to analyze and characterize with specific parameters transient oscillations. These parameters will lead to evaluate physico-chemical conditions as well as the dynamic of the present

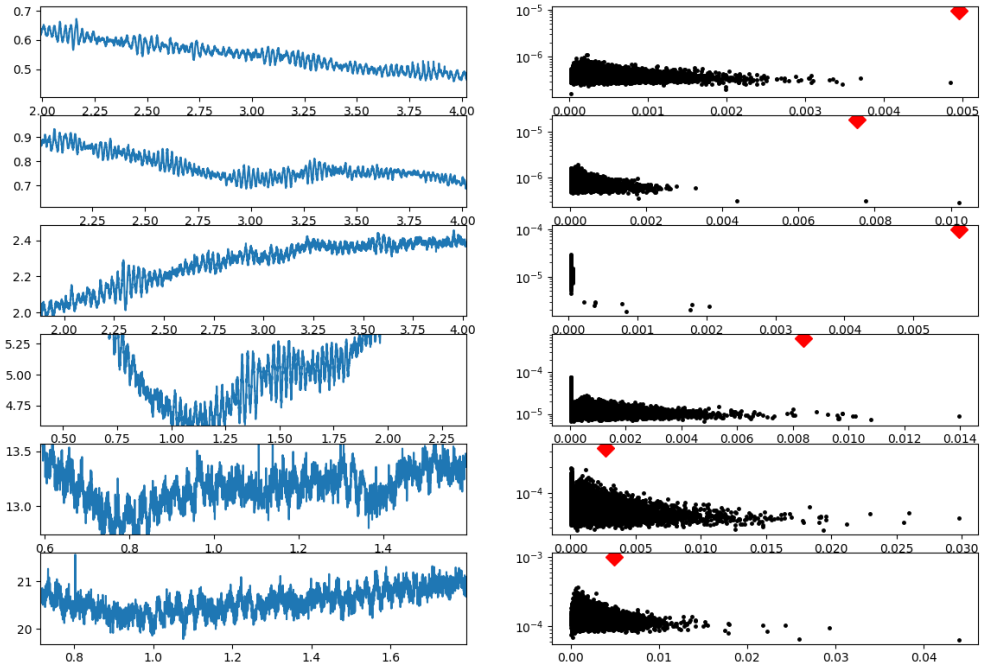

Figure 14: **HF features of the SLS experiments and numerical results of the estimation of the HF features parameters.**(Left column) Zoom on the SLS experimentation signals with initial concentration in μmol from the top to the bottom of $I_0 \in \{0.25,\ 0.35,\ 0.5,\ 1,\ 2,\ ,3\}$ The x-axis is the time in Hours. (Right column) The black dots are the cloud of points $(\widehat{\mathsf{G}}^k_{n,\widehat{m}}, \widehat{\mathsf{D}}^k_{n,\widehat{m}})$ corresponding to the simulation of the null for $k = 1, \ldots, 20000$. The red diamond corresponds to the HF features parameters $(\widehat{\mathsf{G}}_{n,\widehat{m}}, \widehat{\mathsf{D}}_{n,\widehat{m}})$, defined by Definition 1, of the corresponding signal on the left column. The x-axis is the localization parameter of the HF features and the y-axis is the relative amplitude of the HF .

complex system.

5 Conclusion

In this study, we have introduced a method, based on the discrete Fourier transform, to quantify the high frequency features of a given non stationary discrete signal, and then test whether the parameters characterizing these features may be considered as significant or not. We then tested

our method on simulated and experimental data, which shed light on its efficiency, since HF features may be detected even with a noise of the same amplitude. Moreover, the two parameters estimated from the data to characterize the HF are informative *per se*: they could be used by the experimentalists to compare different experimental conditions and their influence on such transient phenomena in the signals. They may also reveal useful in the search for quantitative comparison between mechanistic models, such as the one proposed in [9], and experimental data.

The test to detect HF feature is based on the projection of the signal in a discrete Fourier basis. A further step, in order to localize them, would be to define them in a wavelet basis. The number of parameters will then be equal to three (one for the resolution, one for the amplitude and one for the localisation on the time-scale), and the test of hypothesis has to be extended to this framework. This is a direction for future work.

Acknowledgments. M.D., M.M. and H.R. have been supported by the ERC Starting Grant SKIPPER^{AD} (number 306321).

References

[1] Myles R Allen and Leonard A Smith. Monte carlo ssa: Detecting irregular oscillations in the presence of colored noise. *Journal of climate*, 9(12):3373–3404, 1996.

[2] Leonid Breydo, Natallia Makarava, and Ilia V Baskakov. Methods for conversion of prion protein into amyloid fibrils. In *Prion Protein Protocols*, pages 105–115. Springer, 2008.

[3] David S Broomhead and Roger Jones. Time-series analysis. *Proc. R. Soc. Lond. A*, 423(1864):103–121, 1989.

[4] Jonathan Buckheit, Shaobing Chen, David Donoho, Iain Johnstone, and JD Scargle. Wavelab reference manual. *Version 0.700, December*, 1995.

[5] David L Donoho and Iain M Johnstone. Threshold selection for wavelet shrinkage of noisy data. In *Engineering in Medicine and Biology Society, 1994. Engineering Advances: New Opportunities for Biomedical Engineers. Proceedings of the 16th Annual International Conference of the IEEE*, volume 1, pages A24–A25. IEEE, 1994.

[6] David L Donoho, Iain M Johnstone, et al. Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3):879–921, 1998.

[7] David L Donoho, Iain M Johnstone, Gérard Kerkyacharian, and Dominique Picard. Wavelet shrinkage: asymptopia? *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 301–369, 1995.

[8] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994.

[9] Marie Doumic, Klemens Fellner, Mathieu Mezache, and Human Rezaei. A bi-monomeric, nonlinear Becker-Döring-type system to capture oscillatory aggregation kinetics in prion dynamics. preprint, August 2018.

[10] Nina Golyandina, Vladimir Nekrutkin, and Anatoly A Zhigljavsky. *Analysis of time series structure: SSA and related techniques*. Chapman and Hall/CRC, 2001.

[11] John Edward Jones. On the determination of molecular fields.—ii. from the equation of state of a gas. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, 106(738):463–477, 1924.

[12] Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM review*, 51(2):339–360, 2009.

[13] Giuseppe Legname, Ilia V Baskakov, Hoang-Oanh B Nguyen, Detlev Riesner, Fred E Cohen, Stephen J DeArmond, and Stanley B Prusiner. Synthetic mammalian prions. *Science*, 305(5684):673–676, 2004.

[14] Jiali Li, Shawn Browning, Sukhvir P Mahal, Anja M Oelschlegel, and Charles Weissmann. Darwinian evolution of prions in cell culture. *Science*, 327(5967):869–872, 2010.

[15] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[16] Milan Palus and Dagmar Novotná. Detecting modes with nontrivial dynamics embedded in colored noise: Enhanced monte carlo ssa and the case of climate oscillations. *Physics Letters A*, 248(2-4):191–202, 1998.

[17] Milan Paluš and Dagmar Novotná. Detecting oscillations hidden in noise: Common cycles in atmospheric, geomagnetic and solar data. In *Nonlinear Time Series Analysis in the Geosciences*, pages 327–353. Springer, 2008.

[18] Stéphanie Prigent, Annabelle Ballesta, Frédérique Charles, Natacha Lenuzza, Pierre Gabriel, Léon Matar Tine, Human Rezaei, and Marie Doumic. An efficient kinetic model for assemblies of amyloid fibrils and its application to polyglutamine aggregation. *PLoS One*, 7(11):e43273, 2012.

[19] Stanley B Prusiner. Prions. *Proceedings of the National Academy of Sciences*, 95(23):13363–13383, 1998.

[20] Morten O Ravn and Harald Uhlig. On adjusting the hodrick-prescott filter for the frequency of observations. *Review of economics and statistics*, 84(2):371–376, 2002.

[21] Christian H Reinsch. Smoothing by spline functions. *Numerische mathematik*, 10(3):177–183, 1967.

[22] Robert Vautard, Pascal Yiou, and Michael Ghil. Singular-spectrum analysis: A toolkit for short, noisy chaotic signals. *Physica D: Nonlinear Phenomena*, 58(1-4):95–126, 1992.

[23] Limei Xu, Plamen Ch Ivanov, Kun Hu, Zhi Chen, Anna Carbone, and H Eugene Stanley. Quantifying signals with power-law correlations: A comparative study of detrended fluctuation analysis and detrended moving average techniques. *Physical Review E*, 71(5):051101, 2005.

[24] Antoni Zygmund. *Trigonometric series*, volume 1. Cambridge university press, 2002.

A Materials and methods of the depolymerisation experiment shown in Figures 1 and 12

Formation of amyloid fibrils: PrP amyloid fibrils were formed using the manual setup protocol described previously in [2]. Fibril formation was monitored using a ThT binding assay [2]. Samples were dialysed in 10 mM sodium acetate, pH 5.0. Then fibrils were collected by ultracentrifugation and resuspended in 10 mM sodium acetate, pH 5.0. A washing step was performed by repeating the ultracentrifugation and resuspension steps in 10 mM sodium acetate, pH 5.0. Static light scattering: Static light scattering kinetic experiments were performed with a thermostatic homemade device using a 407-nm laser beam. Light-scattered signals were recorded at a 112^0 angle. Signals were processed with a homemade MatLab program. All experiments have been performed at 55^0C in a 2mmX10mm cuve.

B Library in python to implement the numerical simulation

The numerical simulations have been made with the Python library accessible at *https://github.com/mmezache/HFFTest*. The functions of the library are explicitely commented in the file "README.md". The functions are organized in four categories in the library:

1. the procedure to compute the HF features parameters,
2. the procedure to simulate the null hypothesis,
3. the Monte Carlo procedure to compute the p-value,
4. the procedure to compute test signals such as the ones displayed in Figures 2, 4, 7.

The file "ExampleHFF.py" is a python program which computes the complete procedure for a test signal. The users may change at will the following parameters:

- the length of the signal,
- the standard deviation of the noise,
- the amplitude of the oscillations,
- the parameter of the ℓ^1-trend filtering,
- the number of iteration of the Monte Carlo procedure,
- the choice of the test signal.

The program displays the test signal obtained, the trend estimate, the cloud of points corresponding to the HF features of the null (blue dots) and the point corresponding to the HF features of the tested signal (red dot), and the single-sided amplitude spectrum of the signal which emphasizes the points where the computations of the HF features are performed (cf Figure 3).

The computational time may be significantly long if the number of iterations of the Monte Carlo procedure is large (over 100). However the Monte Carlo procedure can be computed in a parallelized framework which reduces drastically the computational time.

Moreover the automatic choice of the smoothing parameter $\widehat{m}$ is efficient for signals which display oscillations of "high" frequency, i.e. if the spike corresponding to the oscillations in the single-sided amplitude spectrum is located away from the low-frequency components (cf Section 3 and Example 2 in "ExampleHFF.py"). The procedure was designed to identify oscillations "hidden" in the noise, a situation which corresponds to the experimental signals. If the signal tested has oscillations located in the low frequencies, the users are advised to fix the smoothing parameters (see Example 1 in "ExampleHFF.py").