Time Signal Anomaly Detection methodology based on Kernel density estimator and entropic functionals

PhD Candidate: Antonio Squicciarini

III Workshop Junior Interdisciplinar UNED - 31/01/2024

Email: a.squicciarini@alumnos.upm.es
University: Universidad Politécnica de Madrid - UPM ETSII
Department: Departamento de Matemática Aplicada a la Ingeniería Industrial DMAII
Research Group: Teoría de Aproximación Constructiva y Aplicaciones GI TACA
IMEIO program - Ingeniería Matemática, Estadística e Investigación Operativa

Prof. Elio Valero Toranzo[a], Prof. Alejandro Zarzo Altarejos[b], Prof. Carlos E. González Guillén[b]

a)Departamento de Matemática Aplicada, Escuela Superior de Ciencias Experimentales y Tecnología, Universidad Rey Juan Carlos, España, (elio.vtoranzo@urjc.es)

b)GI-TACA, Departamento de Matemática Aplicada a la Ingeniería Industrial, Escuela Técnica Superior de Ingenieros Industriales, Universidad Politécnica de Madrid, España, (alejandro.zarzo@upm.es)

Outline

Outline

In the realm of data-driven decision-making, time series analysis plays a crucial role. Detecting unusual patterns in time series data is vital for assessing the reliability of various systems, ranging from industrial processes and financial markets to healthcare and cybersecurity.

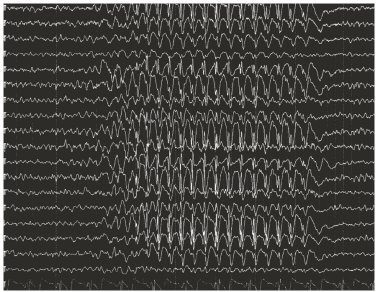


Figure: EEG multichannel seizure
https://commons.wikimedia.org/wiki/File:EEG_Absence_seizure.png

Anomaly Definition

A deviation from "normal behavior" in a time series (Geiger et al. 2020).

- **Point anomaly**: a single data point that has reached an unusual value.
- **Collective anomaly**: a continuous sequence of data points that are considered anomalous as a whole, even if the individual data points may not be unusual.

For collective anomalies, no prior assumption is made regarding the length of the anomalies.

While there is no universal definition of an anomaly in time series, many fields relate it to shifts in the frequency content.

- For instance, in **machine fault detection**, an anomaly can result from a change in machine stiffness, affecting its modal response (Bently et al. 2003).
- Variations in specific frequency bands in EEG signals provide crucial information for detecting **epileptic seizures** (Rosso et al. 2006).
- **Earthquake** frequency data involves electromagnetic emissions (kHz to MHz) from opening cracks, serving as precursors to general fractures. Notably, pre-fracture MHz Electromagnetic radiation precedes kHz signals in both laboratory and geophysical settings (Kalimeri et al. 2008).

Information Theory (Cover and Thomas 2006): branch of applied mathematics and signal processing dealing with information representation, communication, and processing. It produces feature extraction solutions to describe the probability density function (PDF) associated with a random variable X.

- **Entropic measures**: designed to quantify the amount of uncertainty or randomness in a set of data (e.g., Shannon, Tsallis, Rényi entropy).
- **Information measures**: such as non-parametric Fisher information, interpreted as a measure of order/organization of a PDF, complementary to entropic measures.

- Biomedical signals (**EEG**, **ECG**)(Rosso et al. 2006; Eftaxias et al. 2011; Bezerianos, Tong, and Thakor 2003; M. T. Martin, Pennini, and Plastino 1999; M. Martin, A.R. Plastino, and A. Plastino 2000; Zhang, Yang, and Huang 2008; Farashi 2016).
- **Seismic signals** (study of precursor factors) (Eftaxias et al. 2011; Telesca, Lovallo, et al. 2013; Telesca, Chamoli, et al. 2015; Kalimeri et al. 2008).
- **Climatic data** (time signal analysis) (Guignard et al. 2020; Lovallo et al. 2013).

Time-dependent entropy (TDE)

Locate entropy estimation inside a time signal, executed through an overlapping/not overlapping window division strategy: (Farashi 2016; Rosso et al. 2006; Bezerianos, Tong, and Thakor 2003; Kalimeri et al. 2008; M. Martin, A.R. Plastino, and A. Plastino 2000; M. T. Martin, Pennini, and Plastino 1999; Eftaxias et al. 2011; Kalimeri et al. 2008)

- Non-parametric discrete inference or histogram (Farashi 2016; M. T. Martin, Pennini, and Plastino 1999; M. Martin, A.R. Plastino, and A. Plastino 2000; Bezerianos, Tong, and Thakor 2003).
- Symbolic dynamics (Kalimeri et al. 2008; Eftaxias et al. 2011).
- Time-frequency transformation
  - Power spectral density (Zhang, Yang, and Huang 2008).
  - Wavelet entropy (Rosso et al. 2006; Ocak 2009)
- Embedding solutions
  - Approximate entropy (Ocak 2009).
  - Sample and fuzzy entropy (Cao and C.-T. Lin 2018; Cao, Ding, et al. 2020; Xiang et al. 2015; Liang et al. 2015).
  - Kraskov entropy (Patidar and Panigrahi 2017).
- Kernel Density Estimation (KDE) (Guignard et al. 2020).

- **Case-specific application** with little or no focus on main parameter tuning, grouped into two categories:
  - Time-window division parameters
  - Inference parameters
- **Almost exclusively discrete inference**. Limited attention has been given to continuous non-parametric inference solutions.
  - Underlying phenomena are continuous, and the signal is the result of a discretization of it.
  - Some information measures (such as non-parametric Fisher) are not well defined in the discrete case.

Propose a methodology applicable across various domains to apply entropic/information measures to detect anomalies inside time signals, justifying the selection of each method and parameter.

- Utilize **overlapping windows** to obtain time-dependent entropy/information, freeing the window scale from the time step.
- Utilize **different window scales** Δ because the sensitivity to malfunction is highly related to the scale of the window (Farashi 2016).
- Utilize **non-parametric inference** that generates **continuous PDF**
- Adjust the inference parameter based only on a **healthy signal**.
- Inference parameters are highly related to the scale Δ picked. Hence, **optimize the inference for each scale** Δ.

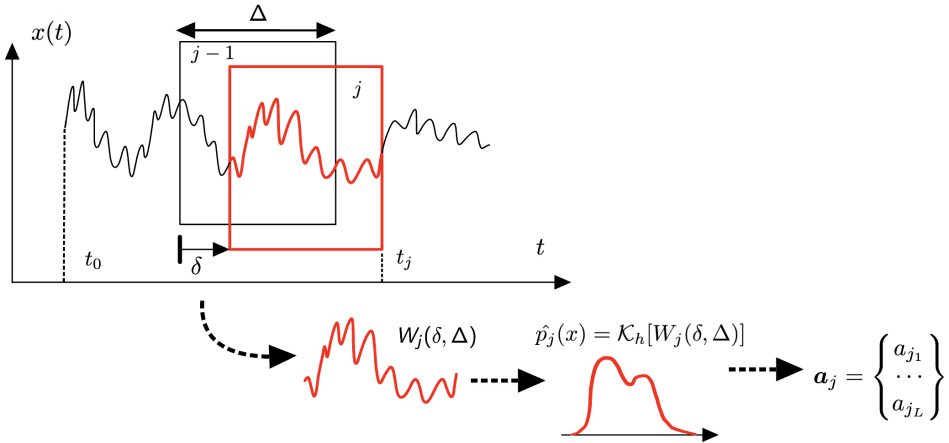


Figure: TDE with one window scale Δ

Outline

X be a continuous random variable
Probability density function (PDF) $p_X : \Lambda \longrightarrow \mathbb{R}^+ \cup \{0\}$ such that $\int_\Lambda p_{(x)}\, dx = 1$

Shannon Entropy

$$S[p] = -\int_\Lambda p(x) \log[p(x)]\, dx \tag{1}$$

Tsallis Entropy

$$T_q[p] = \frac{1}{q-1}\left(1 - \int_\Lambda [p(x)]^q\, dx\right), \quad q \in \mathbb{R} \tag{2}$$

Rényi Entropy

$$R_q[p] = \frac{1}{1-q} \log\left(\int_\Lambda [p(x)]^q\, dx\right), \quad q \in \mathbb{R} \tag{3}$$

- Both in Tsallis and Rényi entropy, the q parameter serves to put more emphasis on the tails ($q < 1$) or on the center of mass ($q > 1$) of the PDF
- Both Rényi and Tsallis entropies reduce to the Shannon entropy (Rényi 1961; Tsallis 1998) in the limit $q \to 1$:

$$\lim_{q \to 1} R_q[p] = S[p] \quad \text{and} \quad \lim_{q \to 1} T_q[p] = S[p],$$

- Shannon and Rényi entropies are **extensive measures**. For iid random variables X and Y, they fulfil the additivity law:

$$S[p_{XY}] = S[p_X] + S[p_Y] \quad \text{and} \quad R_q[p_{XY}] = R_q[p_X] + R_q[p_Y],$$

- Tsallis entropy is non-extensive, and fulfils the pseudo-additivity law:

$$T_q[p_{XY}] = T_q[p_X] + T_q[p_Y] + (1 - q)\, T_q[p_X]\, T_q[p_Y],$$

The complementary information measure considered is the differential Fisher information, introduced by Fisher in 1925 (Cover and Thomas 2006) in the context of statistical estimation:

Parametric Fisher Information (Cover and Thomas 2006)

$$F[p_{X,\theta}] = -\int_{\Lambda} \left[\frac{\partial^2}{\partial \theta^2} \log p(x|\theta) \right] p(x|\theta)\, dx = -\mathbb{E}_{\theta} \left[I''(x|\theta) \right], \tag{4}$$

with $p(x|\theta)$ the PDF associated to a X, θ is an unknown parameter and $I(x|\theta) = \log p(x|\theta)$ the log-likelihood function. Specifically, Fisher information quantifies the information carried by the random variable about the parameter θ.

non-parametric Fisher Information

$$F[p] = \int_{\Lambda} \frac{\left(\frac{d}{dx}p(x)\right)^2}{p(x)}\, dx = \mathbb{E}\left[\left(\frac{\partial}{\partial x}\log p(x)\right)^2\right], \tag{5}$$

When θ is a location parameter (Guignard et al. 2020)

where we assume that $p(x)$ is differentiable and both $p(x)$ and $\frac{d}{dx}p(x)$, are quadratically integrable on $\mathbb{R}$(Bercher 2011).
Fisher information is a non-negative functional which quantifies the average of the proportional change of $p(x)$ per unit change in x. Hence, Fisher information is able to detect the degree of oscillatory character of a given PDF

Shannon Entropy Power

$$N_S[p] = \frac{1}{2\pi e} e^{2S[p]} \tag{6}$$

With $N_S[f] \geq 0$, $\forall\, x \in \Lambda$
In case of Gaussian PDF:

- $N_S[\mathcal{N}(X = x|\mu, \sigma)] = \sigma^2$
- $F[\mathcal{N}(X = x|\mu, \sigma)] = 1/\sigma^2$

Kullback-Leibler

$$KL(p||\rho) = \int_{-\infty}^{\infty} p(x) \ln\left(\frac{p(x)}{\rho(x)}\right) dx \geq 0 \tag{7}$$

A measure of dissimilarity between two probability distributions. It is not symmetric and it is not bounded. (J. Lin 1991).

Jeffrey's Divergence

$$JF(p||\rho) = \frac{1}{2}KL(p||\rho) + \frac{1}{2}KL(\rho||p) \tag{8}$$

To compensate for the asymmetry of Kullback-Leibler

Jensen-Shannon Divergence

$$\begin{aligned} JS^{\pi}(p,\rho) &= \pi_1 KL(p||m) + \pi_2 KL(\rho||m) \\ &= \mathbb{H}[m] - \pi_1 \mathbb{H}[p] - \pi_2 \mathbb{H}[\rho] \end{aligned} \tag{9}$$

with $m(x) = \pi_1 p(x) + \pi_2 \rho(x)$ and $\boldsymbol{\pi}$ is a discrete probability mass function (PMF).

Jensen-Shannon Divergence is symmetric, bounded, and does not require absolute continuity. Generalizable to more than two distributions. (J. Lin 1991)

- Upper-bounded by the entropy of the weight distribution $\boldsymbol{\pi}$.

$$JS^{\pi}(p,\rho) \leq \mathbb{H}[\boldsymbol{\pi}] \tag{10}$$

- Generalized form for multiple PDFs.

$$JS^{\pi}(\{p_j\}_{j=1}^{M}) = \sum_{j=1}^{M} \pi_j KL(p_j||\bar{p})$$
$$= \mathbb{H}\left[\sum_{j=1}^{M} \pi_j p_j\right] - \sum_{j=1}^{M} \pi_j \mathbb{H}[p_j] \quad (11)$$

- Upper bound expressed as the entropy of the PDF weights.

$$JS^{\pi}(\{p_j\}_{j=1}^{M}) \leq \mathbb{H}[\{\pi_j\}_{j=1}^{M}] \quad (12)$$

With $\bar{p}(x) = \sum_{j=1}^{M} \pi_j p_j(x)$

	Symmetric	Bounded	Generalizable ≥ 2 PDFs
Kullback-Leibler	No	No	No
Jeffrey's Divergence	Yes	No	No
Jensen-Shannon Divergence	Yes	Yes	Yes

Table: Caption

Kernel Density Estimation (KDE) (Parzen 1962)

$$\hat{p}_h(x) = \mathcal{K}_h[\{x_i\}_{i=1}^n] = \frac{1}{hn} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{13}$$

Non-parametric inference method that returns a continuous PDF. KDE estimates the PDF at a specific point by summing up the contributions of data points in its vicinity

- $K(x)$ is the kernel, and h is the kernel's bandwidth.
- Kernel, $K(x)$, is assumed to be an even regular function, with unit variance and zero mean.
- Several kernels are available; in our case, we use the Gaussian one.
- Bandwidth selection is more important than kernel selection

For more details, refer to (Parzen 1962; Raykar and Duraiswami 2006; Zambom and Dias 2013).

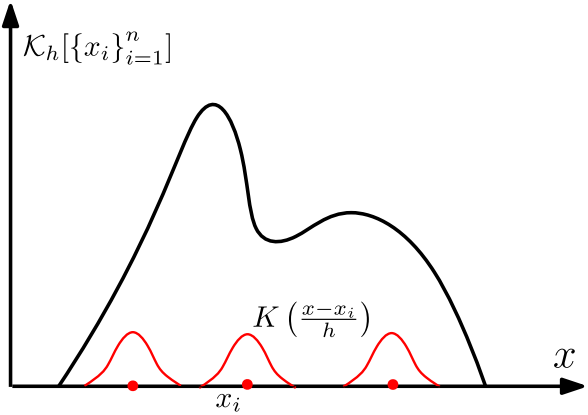


Figure: KDE Representation

The successful application of the KDE method relies on an appropriate selection of the bandwidth h (smoothing parameter) (Raykar and Duraiswami 2006).

- **Underestimated** bandwidth (small h) leads to small bias and large variances (e.g., overfitting).
- **Overestimated** bandwidth (large h) leads to an increase in bias and small variances (e.g., underfitting).

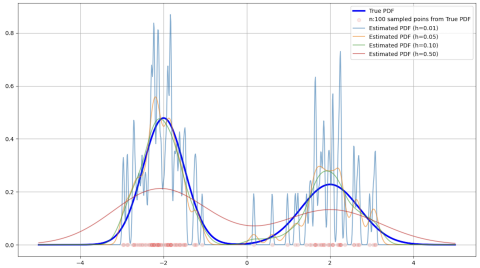


Figure: Applying KDE with different bandwidths (h) to a known PDF (blue line) using finite samplings (red points).

$$\text{MISE}(h) = \mathbb{E}\left[\int_{-\infty}^{+\infty} |p(x) - \hat{p}_h(x)|^2 \, dx\right], \tag{14}$$

The real density $p(x)$ is unknown. Based on a certain assumption, the MISE can be rewite using the Taylor series be expansion after decomposing the MSE into the variance and bias terms:

$$\text{MISE}(h) = \text{AMISE}(h) + o\left(\frac{1}{nh} + h^5\right) \tag{15}$$

Where:

$$\text{AMISE}(h) = \frac{1}{Nh} R(K) + \frac{1}{4} h^4 \mu_2(K)^2 R\left(p''\right), \tag{16}$$

And

$$R(g) = \int_{\mathbf{R}} g(x)^2 dx, \quad , \mu_2(g) = \int_{\mathbf{R}} x^2 g(x) dx \tag{17}$$

Hence, the h optimal is:

$$h = \left(\frac{R(K)}{n\sigma_K^4 R(p'')} \right)^{1/5} \tag{18}$$

AMISE assumption (Raykar and Duraiswami 2006).

- $p''(x)$ is continuous, square-integrable and ultimately monotone
- independently and identically distributed (iid) data

With the AMISE, different h optimization solutions have been proposed. After simplistic assumptions, Silverman (19) and Scott (20) rules are well-known heuristic rules.

$$h = 1.06 \times \min(\hat{\sigma}, \frac{IQR}{1.34}) \times n^{-1/5} \tag{19}$$

$$h = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} \tag{20}$$

Apart from that, more complex solutions are based on plug-in solutions, substituting the real p'' with its approximated version $\hat{p}''$ (Sheather and Jones 1991).

In (Harvey and Oryshchenko 2012), the authors integrate Kernel Density Estimation (KDE) with weighted schemes ω from time series analysis. They determine the parameters that optimize maximum likelihood (21) for filtering or likelihood cross-validation (22) for smoothing applications

$$\ell(\omega, h) = \frac{1}{T-m} \sum_{t=m}^{T-1} \ln \hat{f}_{t+1|t}\left(y_{t+1}\right) = \frac{1}{T-m} \sum_{t=m}^{T-1} \ln \left[\frac{1}{h} \sum_{i=1}^{t} K\left(\frac{y_{t+1} - y_i}{h} \right) w_{t,i}(\omega) \right], \tag{21}$$

$$CV(\omega, h) = \frac{1}{T} \sum_{t=1}^{T} \ln \widehat{f}_{(-t)|T}(y_t) = \frac{1}{T} \sum_{t=1}^{T} \ln \left[\frac{1}{h} \sum_{\substack{i=1 \\ i \neq t}}^{T} K\left(\frac{y_t - y_i}{h} \right) w_{t,T,i}(\omega) \right], \tag{22}$$

With $w_{t,i}(\omega)$ representing a one-sided filter and $w_{t,T,i}(\omega)$ denoting a two-sided smoothing filter

(Garcin 2023) proposed a new complexity metric to balance the overfitting and underfitting of KDE, to apply then the non-parametric inference solution to describe financial time data.

$$\mathcal{C}_h = \min\left(\frac{\mathcal{E}_h}{\max_{\eta\in(0,h_p]}\mathcal{E}_\eta}, \frac{\mathcal{P}_h}{\max_{\eta\in(0,h_p]}\mathcal{P}_\eta}\right) \tag{23}$$

$$\mathcal{E}_h = \sup_{x\in\mathbb{R}}\left|\widehat{F}_h(x) - \frac{1}{n}\sum_{i=1}^{n}\mathbb{L}x_i \leq x\right| \tag{24}$$

$$\mathcal{P}_h = \int_{\mathbb{R}}\widehat{F}_h(x)\log\left(\frac{\widehat{F}_h(x)}{G_{\widehat{\theta}}(x)}\right)dx, \tag{25}$$

Previous h optimization solutions are not specifically designed for time series anomaly detection.

- While there exist instance-specific algorithms, in this application context, it is preferable to have a solution that can be **optimized offline** to enable online monitoring of the system.
- The AMISE assumptions remain unfulfilled with:
  - Non-iid data, as inherent temporal dependencies and patterns may not be removable.
  - The assumption that the underlying probability density function (PDF) is a single-modal distribution (or its equivalent, where the second derivative of the real PDF is monotone) cannot be made.

Outline

Methodology Flowchart

1. Select a finite set of windows $\boldsymbol{\Delta}$
2. For each scale Δ, transform signal in multiple sequences of **ordered** probability density functions (PDFs)
   1. Splitting the signal into an ordered sequence of samples
   2. Transform each sample into a PDF
3. Over the PDFs compute the entropic/information measures

The results are synchronous time-dependent entropic/information sequences. The sequences are synchronous despite having different scale Δ, because all share the same time step δ

Overlapping Windows definition

Sliding temporal window: $W_j(\delta, \Delta) = \{x_i,\ i = 1 + j\delta - \Delta, \cdots, j\delta\}$

- $\Delta \in \mathbb{N}$, $\Delta \leqslant N$ is the window length and $\delta \in \mathbb{N}$, $\delta \leqslant \Delta$ is the sliding factor
- The subscript, $j \in \left[0, \frac{N-\Delta}{\delta}\right] \cap \mathbb{N}$, refers to the time order of the windows
- Window time reference $\tau_j = t_0 + \frac{j\delta}{f_s}$
- Each window is individually standardize $\tilde{x}_i$

Window Labelling

$y_j = 1$ if $t_b < t_j < t_f + \frac{\Delta}{f_s}$

- The timestamps t_b and t_e indicate the beginning/end of the anomaly

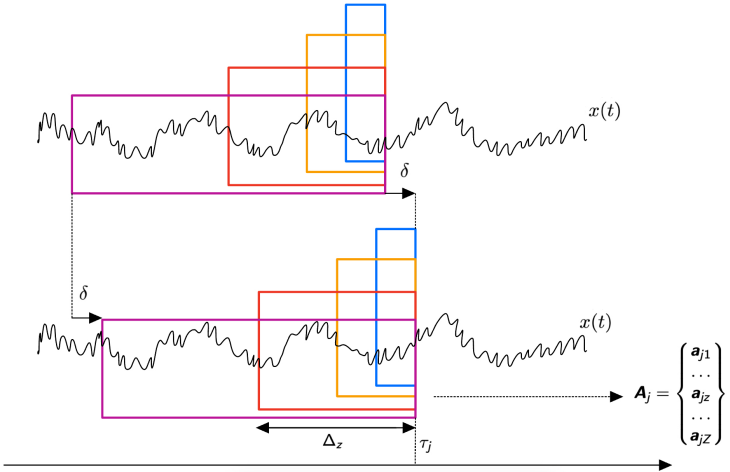


Figure: Syncronous multiscale windows representation

Utilise JSD to control bandwidth optimization, to strike a balance between overfitting and underfitting. Select a h^* for each Δ in the predefined list $\boldsymbol{\Delta}$

JSD Score

$$S^{(JS)}(h, \Delta, \delta) = JS^{\pi}\left[\{\mathcal{K}_h[W_j(\delta, \Delta)]\}_{y_j=0}\right] = JS^{\pi}\left[\left\{\frac{1}{h\Delta}\sum_{i=1+j\delta-\Delta}^{\delta\cdot j} K\left(\frac{x - x_i}{h}\right)\right\}_{y_j=0}\right] \tag{26}$$

With the total number of healthy PDFs equal to M^*

Considering uniform weighting $\pi_j = \frac{1}{M^*} \quad \forall j$, this makes the maximum value of the JSD equal to $\log M^*$

Fixing δ, the JSD score (26) is a monotonic decreasing function, thus the h^* can be determined with optimization techniques such as the bisection method or Newton's method

JSD score - h selection

$$S^{(JS)}(h, \Delta) = th^{JS} * \log M^* \rightarrow h^*(\Delta) \tag{27}$$

The JSD has been selected with respect to the other information divergence metrics because:

- It is bounded, allowing to definition of a threshold proportional concerning the max value
- It accepts more than two distributions as input

The optimization of the th^{JS} can be achieved through a successive application of cross-validation, specifically tailored to the chosen classification algorithm.

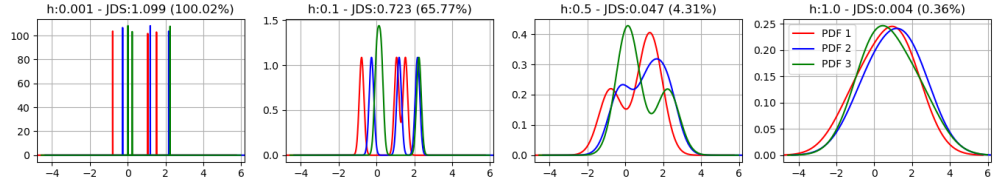


Figure: Representation of how the Jensen-Shannon Divergence (JSD) changes in a simple case with three probability density functions (PDFs), ranging from extremely low to extremely high values of h.

For the bandwidth that tends to 0, the kernel density estimation of the data $\{x_i\}_{i=1}^n$, with $x_i \in \mathbb{R}$ for all i, collapses to the empirical density function:

$$\lim_{h \to 0} \hat{p}(x) = \frac{1}{hn} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) = \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \tag{28}$$

Here, $\delta(x)$ is the Dirac function.

$$\mathbb{H}[\hat{p}(x)] = -\int_{-\infty}^{\infty} \hat{p}(x) \log \hat{p}(x)\, dx = -\int_{-\infty}^{\infty} \frac{1}{n} \sum_{i=1}^{n} \delta(x - x_i) \log \hat{p}(x)\, dx = \tag{29}$$

$$= -\frac{1}{n} \sum_{i=1}^{n} \log \hat{p}(x_i) = -\frac{1}{n} \log \left(\prod_{i=1}^{n} \hat{p}(x_i) \right) \tag{30}$$

Assuming $0 \log 0 = 0$, if the collection $\{x_i\}_{i=1}^n$ is a set (with only unique elements), $\hat{p}(x_i) = \frac{1}{n}$ for all i. In this specific case, $\mathbb{H}[\hat{p}(x)] = -\frac{1}{n} \log \left(\frac{1}{n}\right)^n = \log n$.

Considering two distinct empirical density functions $\hat{p}^{(1)}(x)$ and $\hat{p}^{(2)}(x)$:, generated by sets $\{x_i^{(1)}\}_{i=1}^n$ and $\{x_i^{(2)}\}_{i=1}^n$ each containing the same number of elements and the two sets do not share any points:

$$JS[\hat{p}^{(1)}(x), \hat{p}^{(2)}(x)] = \mathbb{H}[\bar{\hat{p}}(x)] - \frac{1}{2}\mathbb{H}[\hat{p}^{(1)}(x)] - \frac{1}{2}\mathbb{H}[\hat{p}^{(2)}(x)] \tag{31}$$

$$= \log 2n - \frac{1}{2}\log n - \frac{1}{2}\log n = \log(2) \tag{32}$$

Here, $\bar{\hat{p}}(x) = \frac{\hat{p}^{(1)}(x)+\hat{p}^{(2)}(x)}{2}$ is the empirical distribution of $\{x_i^{(1)}\}_{i=1}^n \cup \{x_i^{(2)}\}_{i=1}^n$. Generalizing this result, $JS[\{\hat{p}^{(j)}(x)\}_{j=1}^M] = \log(M)$. In case the sets share some points, $JS[\{\hat{p}^{(j)}(x)\}_{j=1}^M] < \log(M)$.

Instead, consider the opposite case, with $h \to +\infty$, the estimate retains the shape of the used kernel, centred on the mean of the samples (completely smooth).

$$\lim_{h \to \infty} \hat{p}(x) = \mathcal{N}(\mathbb{E}(\{x_i\}_{i=1+j\delta-\Delta}^{\delta \cdot j}), h) \tag{33}$$

Hence, the minimum value of the JSD in this case is equal to:

$$JS^{\pi}\left[\left\{\mathcal{N}(\mathbb{E}(\{x_i\}_{i=1+j\delta-\Delta}^{\delta \cdot j}), h)\right\}_{y_j=0}\right] \tag{34}$$

Considerations:

- δ serves as a "resolution parameter." Its minimum value is set at 1, ensuring that the feature extraction succession matches the frequency of the time signal. However, increasing its value results in a reduction of computational costs.

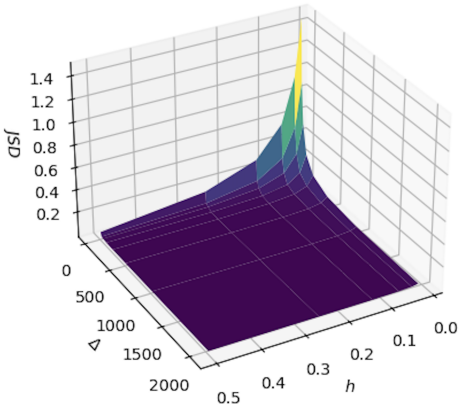


Figure: Representation of the JSD score with respect $\Delta - h$

- Transform each time window into a PDF $\hat{p}_j(x) = \{\mathcal{K}_{h^*}[W_j(\delta, \Delta)]\}_{j=1}^{M}$.
- Associate each PDF with a time-dependent entropic or information feature.
- Create an ordered sequence of matrixes $\{\boldsymbol{A}_1, \boldsymbol{A}_2, \ldots, \boldsymbol{A}_M\}$ with $\boldsymbol{A}_j \in \mathbb{R}^{Z \times L}$
- Each $\boldsymbol{a}_j \in \mathbb{R}^L$ contains L different entropic or information outputs.
- Simultaneous use of various metrics reveals unique characteristics.

$$\boldsymbol{A}_j = \begin{Bmatrix} \boldsymbol{a}_{j1} \\ \ldots \\ \boldsymbol{a}_{jz} \\ \ldots \\ \boldsymbol{a}_{jZ} \end{Bmatrix} = \begin{Bmatrix} a_{j11} & \ldots & a_{j1l} & \ldots & a_{j1L} \\ \ldots & & & & \\ a_{jz1} & \ldots & a_{jzl} & \ldots & a_{jzL} \\ \ldots & & & & \\ a_{jZ1} & \ldots & a_{jZl} & \ldots & a_{jZL} \end{Bmatrix} \tag{35}$$

Outline

To test the capability of our algorithm in a controlled environment, we conduct synthetic experiments. These experiments are based on a multi-harmonic time signal, where new tones are introduced after the timestamp t_b. The intensity of these tones increases linearly until t_f. The signal can be described as follows:

$$x(t) = g(t) \sum_{k=1}^{K_n} \mathbf{Re}\left(A_k \; e^{-i(2\pi f_k t + \phi_k)}\right) + (1 - g(t)) \sum_{k=1}^{K_a} \mathbf{Re}\left(A_k^{(a)} \; e^{-i\left(2\pi f_k^{(a)} t + \phi_k^{(a)}\right)}\right) + \epsilon \tag{36}$$

where ϵ is a white noise applied to the signal, and $g(t)$ is the modulate function that controls the transition from a healthy signal to an anomaly one.

$$\begin{cases} g(t) = 1 & \text{if } t \leq t_b \\ g(t) = 1 - \frac{t - t_b}{t_f - t_b} & \text{if } t > t_b \end{cases} \tag{37}$$

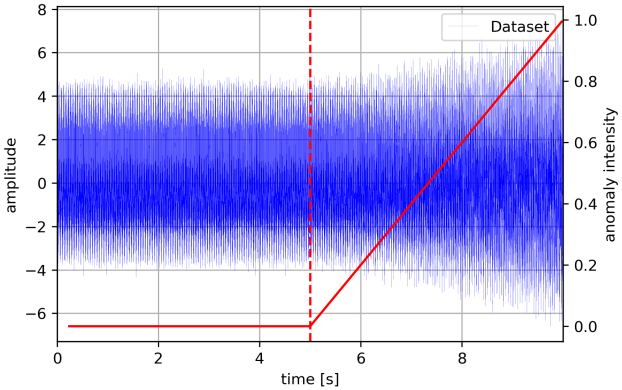


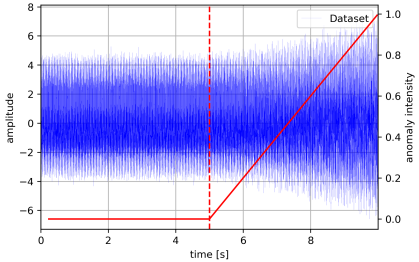
Figure: Synthetic signal representation

sampling_rate	4096
δ	256
$\mathbf{\Delta}$	$2^{[4,5,\ldots,11]}$
th^{JS}	0.001

Table: Main transformation parameters

freqs $[Hz]$	440.0, 220.0, 22.0
amps	1.5, 2.0, 1.0
freqs_anom $[Hz]$	440.0, 220.0, 22.0, 50.0, 1000.0
amps_anom	1.5, 2.0, 1.0, 2.5, 1.0
white_noise_var	0.3
t_0 $[s]$	0.0
t_b $[s]$	5.0
t_f $[s]$	10.0

Table: Synthetic signal parameters



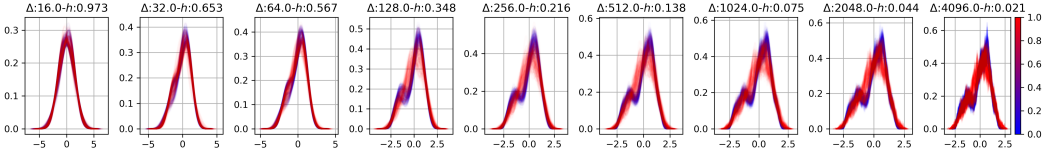


Figure: Represention of all the PDFs generated from the signal at each scale Δ, with its relative selected h. The red colour gradient indicates the anomaly intensity of the PDFs.

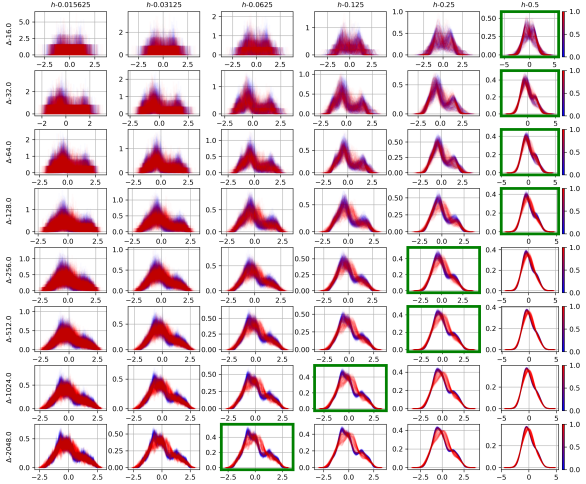


Figure: PDFs comparative grid illustrating the impact of bandwidth h on overfitting and underfitting. Low h aligns closely with empirical data, while high h mirrors the kernel. The solution found by the JSD score criteria is highlighted in green.

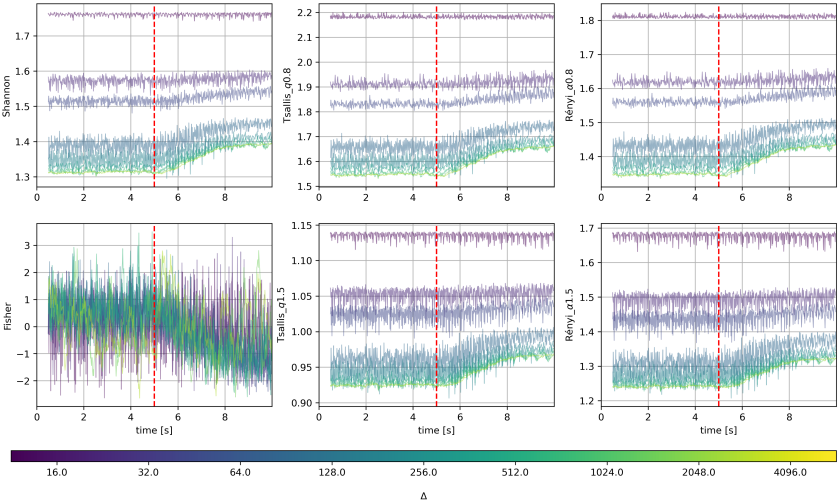


Figure: Entropic and information Time-Dependent plots related to the synthetic experiment. The color gradient indicates the Δ scale of the signal

Outline

Seizure Definition

An epileptic seizure is defined as a transient occurrence of signs and/or symptoms caused by abnormal excessive or synchronous neuronal activity in the brain

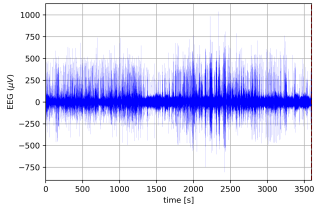


Figure: EEG signal: record chb01-01 Channel 1 (FP1-F7)

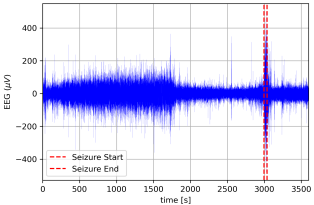


Figure: EEG signal: record chb01-03 Channel 1 (FP1-F7)

Scalp EEG (Electroencephalography)

Record of electrical brain activity using electrodes placed on the scalp. It is a **non-invasive** technique. It is a multichannel signal.

Children's Hospital Boston (CHB-MIT) Scalp EEG database (Shoeb 2010)

- Pediatric subjects with intractable seizures
- $f_z = 256[H_z]$ with a 16-bit resolution signal
- EEG electrode positions follow the international 10-20 system

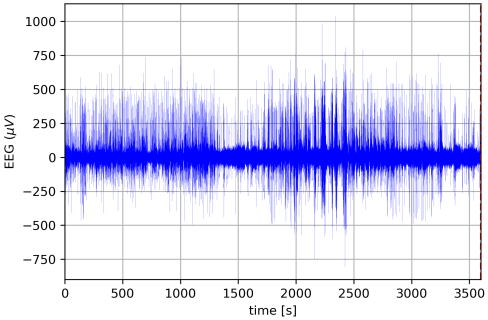


Figure: EEG signal: record chb01-01 Channel 1 (FP1-F7)

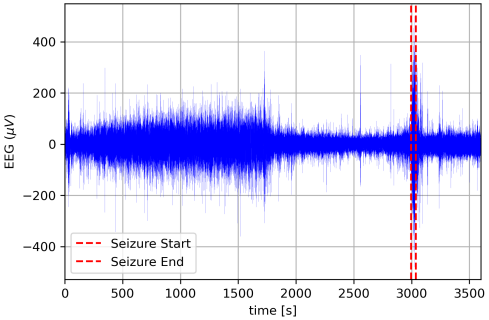


Figure: EEG signal: record chb01-03 Channel 1 (FP1-F7)

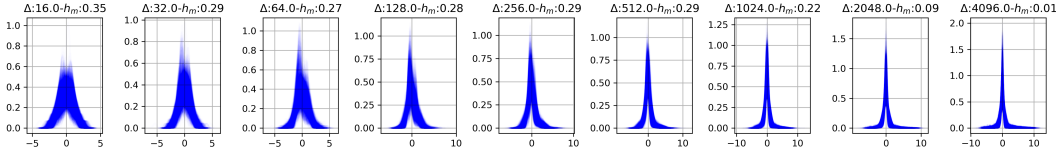


Figure: Rerepresentation of all the PDFs generated from the signal at each scale Δ, with its relative selected h

sampling_rate	256
δ	256
$\mathbf{\Delta}$	$2^{[4,5,\ldots,13]}$
th^{JS}	0.01

Table: chb01: parameters

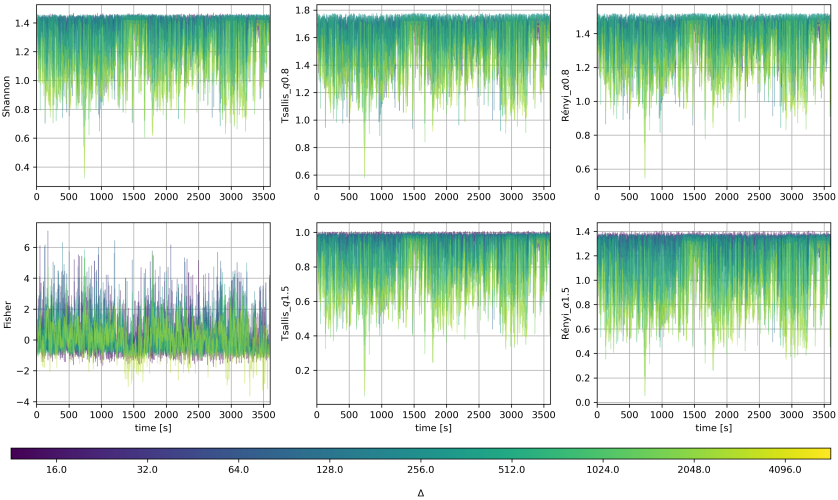


Figure: Entropic and information Time-Dependent plots related to the record chb01-01 Channel 1 (FP1-F7). The colour gradient indicates the Δ scale of the signal.

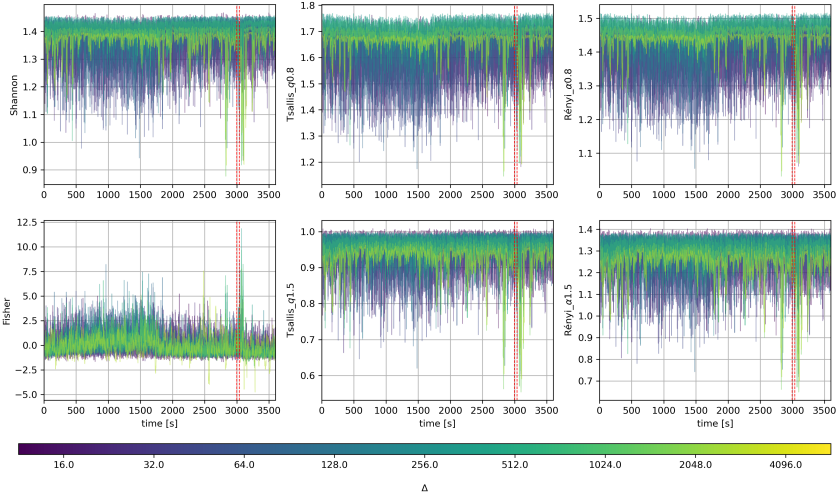


Figure: Entropic and information Time-Dependent plots related to the record chb01-03 Channel 1 (FP1-F7). The colour gradient indicates the Δ scale of the signal.

- Methodology that allows the use of any entropy/informative metric to describe a temporal signal.
- New algorithm for selecting the bandwidth (h) of the KDE designed for time anomaly detection based on the Jensen-Shannon divergence.
- Empirically demonstrating with synthetic and real-world tests that a multiscale solution is crucial for anomaly detection

Thank you!

The work was made possible thanks to the Programa Propio de la Universidad Politécnica de Madrid UPM

Outline

Bently, Donald E. et al. (Dec. 1, 2003). "Fundamentals of Rotating Machinery Diagnostics". In: *Mechanical Engineering-CIME* 125.12, pp. 53–54. ISSN: 00256501.
Bercher, J.-F. (Mar. 14, 2011). "On Escort Distributions, Q-gaussians and Fisher Information". In: *AIP Conference Proceedings* 1305.1, pp. 208–215. ISSN: 0094-243X.
Bezerianos, A., S. Tong, and N. Thakor (Feb. 2003). "Time-Dependent Entropy Estimation of EEG Rhythm Changes Following Brain Ischemia". In: *Annals of Biomedical Engineering* 31.2, pp. 221–232. ISSN: 0090-6964.
Cao, Zehong, Weiping Ding, et al. (May 14, 2020). "Effects of Repetitive SSVEPs on EEG Complexity Using Multiscale Inherent Fuzzy Entropy". In: *Neurocomputing* 389, pp. 198–206. ISSN: 0925-2312.
Cao, Zehong and Chin-Teng Lin (Apr. 2018). "Inherent Fuzzy Entropy for the Improvement of EEG Complexity Evaluation". In: *IEEE Transactions on Fuzzy Systems* 26.2, pp. 1032–1035. ISSN: 1941-0034.
Cover, Thomas M and Joy A Thomas (2006). "ELEMENTS OF INFORMATION THEORY". In: p. 774.
Eftaxias, K. et al. (Oct. 10, 2011). *Are Epileptic Seizures Quakes of the Brain? An Approach by Means of Nonextensive Tsallis Statistics*. arXiv: 1110.2169 [physics]. URL: http://arxiv.org/abs/1110.2169 (visited on 10/12/2022). preprint.
Farashi, Sajjad (Feb. 2016). "A Multiresolution Time-Dependent Entropy Method for QRS Complex Detection". In: *Biomedical Signal Processing and Control* 24, pp. 63–71. ISSN: 17468094.
Garcin, Matthieu (May 22, 2023). *Complexity Measure, Kernel Density Estimation, Bandwidth Selection, and the Efficient Market Hypothesis*. arXiv: 2305.13123 [q-fin, stat]. URL: http://arxiv.org/abs/2305.13123 (visited on 11/05/2023). preprint.
Geiger, Alexander et al. (Dec. 2020). "TadGAN: Time Series Anomaly Detection Using Generative Adversarial Networks". In: *2020 IEEE International Conference on Big Data (Big Data)*. 2020 IEEE International Conference on Big Data (Big Data), pp. 33–43.
Guignard, Fabian et al. (July 14, 2020). "Advanced Analysis of Temporal Data Using Fisher-Shannon Information: Theoretical Development and Application in Geosciences". In: *Frontiers in Earth Science* 8, p. 255. ISSN: 2296-6463.
Harvey, Andrew and Vitaliy Oryshchenko (Jan. 1, 2012). "Kernel Density Estimation for Time Series Data". In: *International Journal of Forecasting*. Special Section 1: The Predictability of Financial Markets 28.1, pp. 3–14. ISSN: 0169-2070.
Kalimeri, M. et al. (Feb. 2008). "Dynamical Complexity Detection in Pre-Seismic Emissions Using Nonadditive Tsallis Entropy". In: *Physica A: Statistical Mechanics and its Applications* 387.5-6, pp. 1161–1172. ISSN: 03784371.
Liang, Zhenhu et al. (2015). "EEG Entropy Measures in Anesthesia". In: *Frontiers in Computational Neuroscience* 9. ISSN: 1662-5188.
Lin, J. (Jan. 1991). "Divergence Measures Based on the Shannon Entropy". In: *IEEE Transactions on Information Theory* 37.1, pp. 145–151. ISSN: 1557-9654.
Lovallo, Michele et al. (Dec. 2013). "Investigating the Time Dynamics of Monthly Rainfall Time Series Observed in Northern Lebanon by Means of the Detrended Fluctuation Analysis and the Fisher-Shannon Method". In: *Acta Geophysica* 61.6, pp. 1538–1555. ISSN: 1895-6572, 1895-7455.
Martin, M. T, F Pennini, and A Plastino (May 31, 1999). "Fisher's Information and the Analysis of Complex Signals". In: *Physics Letters A* 256.2, pp. 173–180. ISSN: 0375-9601.

Martin, M.T., A.R. Plastino, and A. Plastino (Jan. 2000). "Tsallis-like Information Measures and the Analysis of Complex Signals". In: *Physica A: Statistical Mechanics and its Applications* 275.1-2, pp. 262–271. ISSN: 03784371.

Ocak, Hasan (Mar. 1, 2009). "Automatic Detection of Epileptic Seizures in EEG Using Discrete Wavelet Transform and Approximate Entropy". In: *Expert Systems with Applications* 36 (2, Part 1), pp. 2027–2036. ISSN: 0957-4174.

Parzen, Emanuel (Sept. 1962). "On Estimation of a Probability Density Function and Mode". In: *The Annals of Mathematical Statistics* 33.3, pp. 1065–1076. ISSN: 0003-4851, 2168-8990.

Patidar, Shivnarayan and Trilochan Panigrahi (Apr. 1, 2017). "Detection of Epileptic Seizure Using Kraskov Entropy Applied on Tunable-Q Wavelet Transform of EEG Signals". In: *Biomedical Signal Processing and Control* 34, pp. 74–80. ISSN: 1746-8094.

Raykar, Vikas Chandrakant and Ramani Duraiswami (Apr. 20, 2006). "Fast Optimal Bandwidth Selection for Kernel Density Estimation". In: *Proceedings of the 2006 SIAM International Conference on Data Mining (SDM)*. Proceedings. Society for Industrial and Applied Mathematics, pp. 524–528. ISBN: 978-0-89871-611-5.

Rényi, Alfréd (1961). "On Measures of Entropy and Information". In: *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. Vol. 4. University of California Press, pp. 547–562.

Rosso, O.A. et al. (June 2006). "EEG Analysis Using Wavelet-Based Information Tools". In: *Journal of Neuroscience Methods* 153.2, pp. 163–182. ISSN: 01650270.

Sheather, S. J. and M. C. Jones (1991). "A Reliable Data-Based Bandwidth Selection Method for Kernel Density Estimation". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 53.3, pp. 683–690. ISSN: 2517-6161.

Shoeb, Ali (2010). *CHB-MIT Scalp EEG Database*. physionet.org.

Telesca, Luciano, Ashutosh Chamoli, et al. (July 2015). "Investigating the Tsunamigenic Potential of Earthquakes from Analysis of the Informational and Multifractal Properties of Seismograms". In: *Pure and Applied Geophysics* 172.7, pp. 1933–1943. ISSN: 0033-4553, 1420-9136.

Telesca, Luciano, Michele Lovallo, et al. (Aug. 15, 2013). "Fisher–Shannon Analysis of Seismograms of Tsunamigenic and Non-Tsunamigenic Earthquakes". In: *Physica A: Statistical Mechanics and its Applications* 392.16, pp. 3424–3429. ISSN: 0378-4371.

Tsallis, Constantino (Aug. 1, 1998). "Generalized Entropy-Based Criterion for Consistent Testing". In: *Physical Review E* 58.2, pp. 1442–1445.

Xiang, Jie et al. (Mar. 30, 2015). "The Detection of Epileptic Seizure Signals Based on Fuzzy Entropy". In: *Journal of Neuroscience Methods* 243, pp. 18–25. ISSN: 0165-0270.

Zambom, Adriano Z. and Ronaldo Dias (Apr. 1, 2013). "A Review of Kernel Density Estimation with Applications to Econometrics". In: *International Econometric Review* 5.1 (1), pp. 20–42. ISSN: 1308-8793.

Zhang, Aihua, Bin Yang, and Ling Huang (May 2008). "Feature Extraction of EEG Signals Using Power Spectral Entropy". In: *2008 International Conference on BioMedical Engineering and Informatics*. 2008 International Conference on BioMedical Engineering and Informatics. Vol. 2, pp. 435–439.