Adaptive Bandwidth Selection in Nonparametric Kernel Density Estimation for Time Signal Anomaly Detection using Jensen-Shannon Divergence Controlling

PhD Candidate: Antonio Squicciarini

ECM2024 - July 19, 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

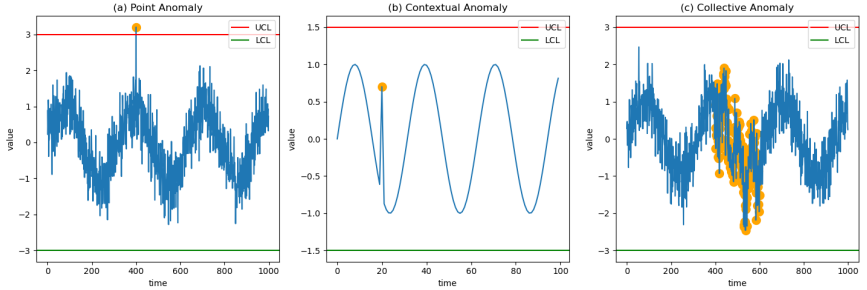

Figure: Anomaly types in time-series data (Choi et al. 2021)

In many fields, it is related to shifts in the frequency content.

- **Machine Fault detection**: an anomaly can result from a change in machine stiffness, affecting its modal response (Bently and Hatch'Charles 2003).
- **Epileptic seizures**: Variations in specific frequency bands in EEG signals provide crucial information for detecting **epileptic seizures** (Rosso et al. 2006).

Locate entropy estimation inside a time signal, executed through an overlapping/not overlapping window division strategy.

- **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.

Applications:

- **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).

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

$$\textbf{Shannon Entropy}: \quad S[p] = -\int_\Lambda p(x) \log[p(x)]\, dx$$

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

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

$$\textbf{Non-parametric Fisher Information}: \quad F[p] = \int_\Lambda \frac{\left(\frac{d}{dx} p(x)\right)^2}{p(x)}\, dx = \mathbb{E}\left[\left(\frac{\partial}{\partial x} \log p(x)\right)^2\right],$$

- 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).

- **Almost exclusively discrete inference**
  - Underlying phenomena are continuous, and the signal results from its discretization.
  - Some information measures (e.g., non-parametric Fisher information) are not uniquely defined in the discrete case.
  - Entropy for continuous random variables is not equal to the entropy limit for discrete random variables for $\Delta \to 0$ the limit diverges.

$$\lim_{\Delta \to 0} \left(H(X^{\Delta}) + \log \Delta \right) = H(X)$$

- **Limited focus on the inference step**
- **Single window scale deployment**

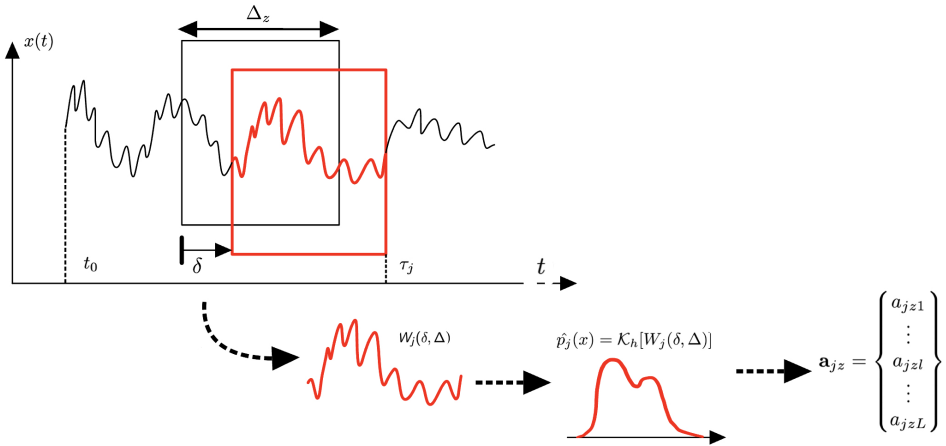

Figure: TDE with one window scale Δ

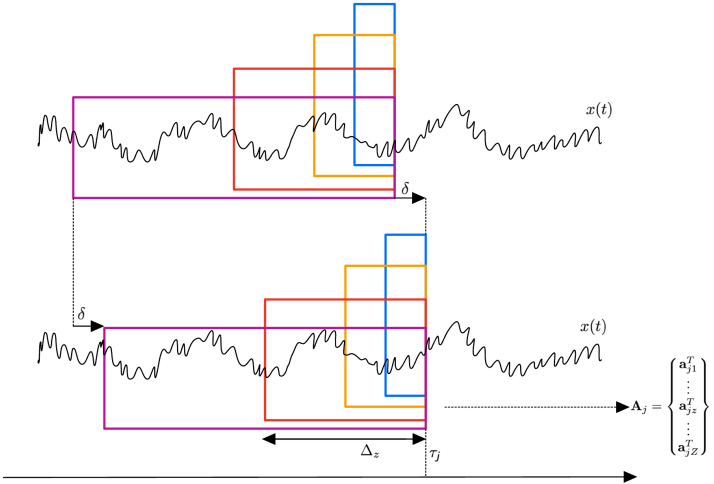

Figure: Syncronous multiscale windows representation

Kernel Density Estimation (KDE): Non-parametric inference method that returns a continuous PDF (Parzen 1962; Raykar and Duraiswami 2006; Zambom and Dias 2013).

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

- $K(x)$ 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.

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

- **Underestimated** bandwidth (small h) leads to small bias and large variance (e.g., overfitting).
- **Overestimated** bandwidth (large h) leads to increased bias and small variance (e.g., underfitting).

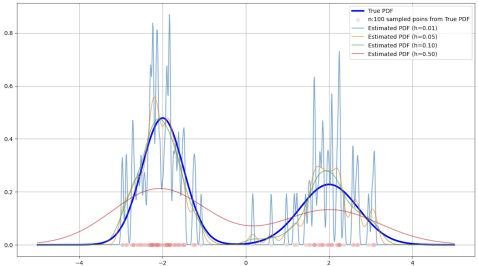

Figure: Applying KDE with different bandwidths (h) to a known PDF (blue line) using finite samples (red points).

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

where

$$R(g) = \int_{\mathbb{R}} g(x)^2 \, dx, \quad \mu_2(g) = \int_{\mathbb{R}} x^2 g(x) \, dx \tag{3}$$

Hence, the optimal h is:

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

Silverman's rule of thumb (5):

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

Complex solutions use plug-in methods, substituting the real p'' with its approximated version $\hat{p}''$ (Sheather and Jones 1991).

AMISE assumptions (Raykar and Duraiswami 2006):

- Data are independently and identically distributed (iid).
- $p''(x)$ is continuous, square-integrable, and ultimately monotone.
- $\lim_{N\to\infty} h = 0$ and $\lim_{N\to\infty} Nh = \infty$, i.e., as the number of samples N is increased h approaches zero at a rate slower than $1/N$.
- $K(x) \geq 0$ and $\int_{\mathbb{R}} K(x)dx = 1$. The kernel function is assumed to be symmetric about the origin ($\int_{\mathbb{R}} xK(x)dx = 0$) and has finite second moment ($\int_{\mathbb{R}} x^2K(x)dx < \infty$).

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

- The AMISE assumptions are unfulfilled when:
  - The data are **non-iid**, as inherent temporal dependencies and patterns may not be removable.
- While instance-specific algorithms exist, in this application context, it is preferable to have a solution that can be **optimized offline** to enable online monitoring of the system.

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 (6) for filtering or likelihood cross-validation (7) for smoothing applications.

$$\ell(\omega, h) = \frac{1}{T-m} \sum_{t=m}^{T-1} \ln \hat{f}_{t+1|t}\left(y_{t+1}\right) = \frac{1}{T-m} \sum_{t=m}^{T-1} \ln \left[\frac{1}{h} \sum_{i=1}^{t} K\left(\frac{y_{t+1} - y_i}{h} \right) w_{t,i}(\omega) \right], \quad (6)$$

$$CV(\omega, h) = \frac{1}{T} \sum_{t=1}^{T} \ln \widehat{f}_{(-t)|T}(y_t) = \frac{1}{T} \sum_{t=1}^{T} \ln \left[\frac{1}{h} \sum_{\substack{i=1 \\ i \neq t}}^{T} K\left(\frac{y_t - y_i}{h} \right) w_{t,T,i}(\omega) \right], \quad (7)$$

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

Kullback-Leibler Divergence

$$KL(p \parallel \rho) = \int_{-\infty}^{\infty} p(x) \ln \left(\frac{p(x)}{\rho(x)} \right) dx \geq 0 \tag{8}$$

Jensen-Shannon Divergence

$$JS^{\pi}(\{p_j\}_{j=1}^{M}) = \sum_{j=1}^{M} \pi_j KL(p_j \parallel \bar{p}) = \mathbb{H}\left[\sum_{j=1}^{M} \pi_j p_j \right] - \sum_{j=1}^{M} \pi_j \mathbb{H}[p_j] \tag{9}$$

where $m(x) = \sum_{j=1}^{M} \pi_j p_j$ and $\boldsymbol{\pi}$ is a discrete probability mass function (PMF).

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

- Upper-bounded: $JS^{\pi}(\{p_j\}_{j=1}^{M}) \leq \mathbb{H}[\{\pi_j\}_{j=1}^{M}]$

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-h Score

$$S^{(JS)}(h, \Delta, \delta) = JS^{\pi}\left[\{\mathcal{K}_h[W_j(\delta, \Delta)]\}_{y_j=0}\right] = JS^{\pi}\left[\left\{\frac{1}{h\Delta}\sum_{i=1+j\delta-\Delta}^{\delta\cdot j} K\left(\frac{x - x_i}{h}\right)\right\}_{y_j=0}\right] \tag{10}$$

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 $\log M^*$

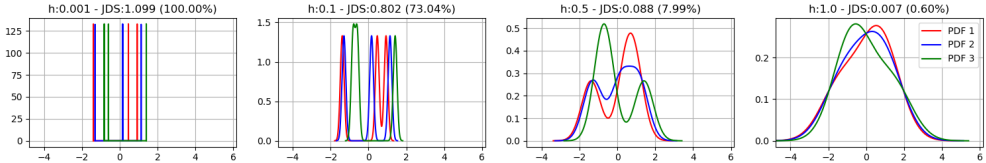

Figure: Representation of how the Jensen-Shannon Divergence (JSD) changes in a simple case with three probability density functions (PDFs), ranging from extremely low to extremely high values of h.

Fixing δ and Δ, the JSD score (10) 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

$$S^{(JS)}(h, \Delta) = th^{JS} * \log M^* \rightarrow h^*(\Delta) \tag{11}$$

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

$$x(t) = g(t) \sum_{k=1}^{K_n} \mathbf{Re}\left(A_k \; e^{-i(2\pi f_k t + \phi_k)}\right) + (1 - g(t)) \sum_{k=1}^{K_a} \mathbf{Re}\left(A_k^{(a)} \; e^{-i\left(2\pi f_k^{(a)} t + \phi_k^{(a)}\right)}\right) + \epsilon(t), \tag{12}$$

Linear increasing anomaly:

$$\begin{cases} g(t) = 1 & \text{if } t \leq t_b \\ g(t) = 1 - \frac{t - t_b}{t_f - t_b} & \text{if } t > t_b \end{cases} \tag{13}$$

Localised contextual anomaly:

$$\begin{cases} g(t) = 1 & \text{if } t \leq t_b \\ g(t) = 1 - \frac{f(t)}{\max(f(t))} & \text{if } t > t_b, \end{cases} \tag{14}$$

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

Table: Main transformation parameters

ϕ_k	440.0, 220.0, 22.0
A_k	1.5, 2.0, 1.0
$\phi_k^{(a)}$	440.0, 220.0, 22.0, 50.0, 1000.0
$A_k^{(a)}$	1.0, 1.0, 0.5, 2.0, 0.5
σ_ϵ	0.5
t_0	0.0
t_b	5.0
t_f	10.0

Table: Synthetic signal parameters. Time in [s] and frequencies in [Hz]

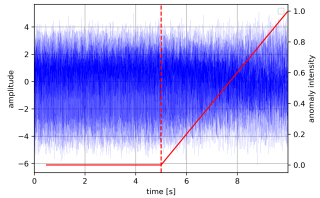

Figure: Example of a synthetic signal, where $g(t)$ is depicted with a continuous red line, representing the **linear increasing anomaly function**.

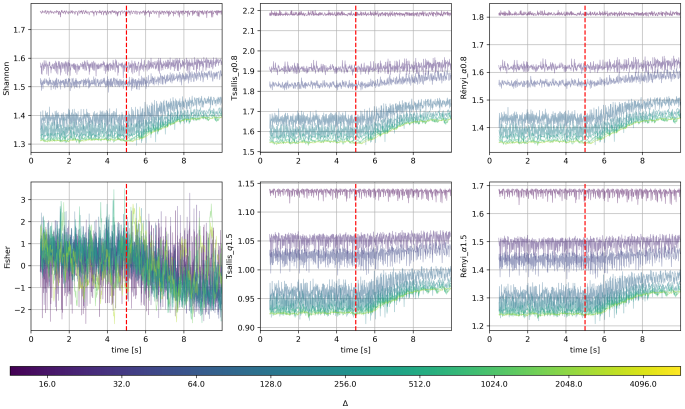

Figure: Entropic and information Time-Dependent plots related to the synthetic experiment. The color gradient indicates the Δ scale of the signal

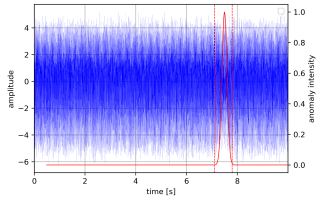

Figure: Example of a synthetic signal, where $g(t)$ is depicted with a continuous red line, representing the **Gaussian localization anomaly function**.

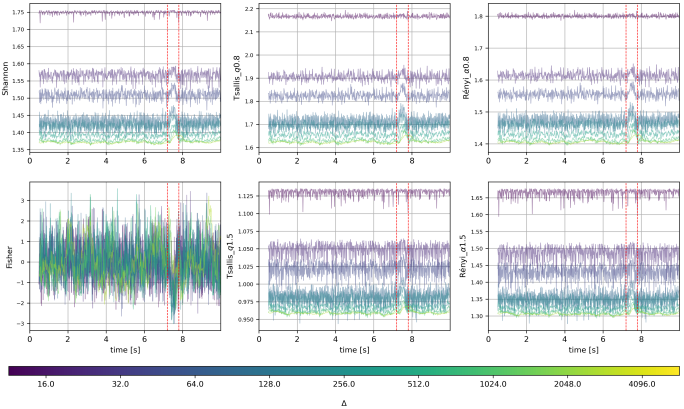

Figure: Entropic and information Time-Dependent plots related to the synthetic experiment. The color gradient indicates the Δ scale of the signal

Thank you!

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

Bently, Donald E and T Hatch'Charles (2003). "Fundamentals of Rotating Machinery Diagnostics". In: *Mechanical Engineering-CIME* 125.12, pp. 53–54.
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 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.
Choi, Kukjin et al. (2021). "Deep Learning for Anomaly Detection in Time-Series Data: Review, Analysis, and Guidelines". In: *IEEE Access* 9, pp. 120043–120065. ISSN: 2169-3536.
Eftaxias, K. et al. (Oct. 2011). *Are Epileptic Seizures Quakes of the Brain? An Approach by Means of Nonextensive Tsallis Statistics*. arXiv: 1110.2169 [physics].
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.
Guignard, Fabian et al. (July 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. 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 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. 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. 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. 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.
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.
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. 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.
Xiang, Jie et al. (Mar. 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. 2013). "A Review of Kernel Density Estimation with Applications to Econometrics". In: *International Econometric Review* 5.1, 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*. Vol. 2, pp. 435–439.