Implementation of a new statistical comparison among signals

Department of Physics and Astronomy, Bologna

Pattern Recognition Exam

Chiara Malvaso

February, 2025

Contents

1 Introduction

In real-world scenarios, signals are inevitably affected by noise. While extensive research has been dedicated to noise reduction, this study focused on assessing whether two signals, contaminated by different noise levels, can be considered equivalent based on their spectral characteristics. Specifically, the aim was to develop a statistical test that determines, with a given confidence level, whether two signals should be classified as identical or distinct.

In this context, signal equivalence is defined in terms of frequency content: two signals are considered the same if they exhibit identical frequency content, regardless of the superimposed noise. Ideally, signals differing only by noise should be identified as equivalent, whereas those with distinct spectral compositions should be classified as different.

To achieve this, the analysis was conducted on a controlled dataset of simulated signals, allowing a systematic evaluation of statistical methodologies. The objective was to establish a statistical test that, given any two signals, applies a rigorous procedure to determine their equivalence, providing a decision with an associated p-value to quantify confidence.

Among various noise types, white noise represents a fundamental and commonly encountered case. Consequently, it is the primary focus of this study, where the goal was to assess whether two signals, each affected by different levels of white noise (or slight variations thereof, see 2), correspond to the same underlying signal or distinct ones. This analysis served as a foundation for extending the methodology to more complex noise models and real-world scenarios.

All the processing and analysis conducted in this study were implemented in Python, and the associated code is available in the following GitHub repository:

https://github.com/malvasochiara/statistical-signal-comparison

The repository contains two folders:

- *main_classification_pipeline*: This folder includes the functions for dataset generation and preparation for statistical analysis, as detailed in sections 2 and 3, as well as the statistical methodology outlined in 4.3.

- *supplementary_material*: This folder contains additional approaches and analyses referenced throughout the report.

1.1 White Noise Properties

White Gaussian Noise (WGN) is one of the most commonly employed stochastic models in practical applications. In signal processing, the term "white noise" traditionally refers to a stochastic process consisting of independent random variables. The origin of this terminology lies in the spectral properties of these stochastic processes, which exhibit a flat spectrum, meaning that all frequencies have equal power.

A stochastic process $\mathbf{X}(t)$ is defined as White Gaussian Noise if it is normally distributed and its values at different time instants, $\mathbf{X}(t_1)$ and $\mathbf{X}(t_2)$, are statistically independent for $t_1 \neq t_2$.The first condition ensures the "Gaussian" nature of the noise, while the second justifies the term "white." Formally,

$$\mathbf{X}(t) \sim \mathcal{N}(0, \sigma^2)$$

which implies that $\mathbf{X}(t)$ has zero mean and a finite variance σ^2. For a discrete-time WGN process, this variance is directly related to the power spectral density.

A key property of WGN is that its Power Spectral Density (PSD) remains constant across all frequencies [1]. This constant value is typically denoted by N_0, leading to the following expression:

$$PSD = \frac{N_0}{2}, \quad \forall f.$$

It is crucial to emphasize that WGN does not accurately describe any real physical phenomenon. By definition, spectrally white noise has a constant spectral density across all frequencies. However, real noise is never truly white but rather remains approximately constant up to a cutoff frequency, beyond which it decreases to ensure a finite variance.

Nevertheless in many practical cases, for example where the Central Limit Theorem applies, WGN remains a valid and useful approximation [2]. Consequently, it serves as a widely used model for generating processes that approximate real physical phenomena [3].

In numerical simulations, sequences of normally distributed random numbers are often employed as an approximation of a WGN process.

2 Dataset construction

The initial phase of the study involved the construction of a suitable dataset to address the problem at hand. The approach was to maintain a general framework, minimizing unnecessary complexities, particularly in the preliminary stages of the analysis. Guided by Fourier's theorem, which asserts that any periodic signal can be represented as a sum of sine and cosine functions, the dataset was constructed as a superposition of sinusoidal waves. This choice ensured a straightforward yet robust foundation for the subsequent analysis. Each signal $S(t)$ is represented as a sum of sinusoidal components, as given by the following equation:

$$S(t) = \sum_{n=1}^{N} \sin(2\pi f_n t) \tag{1}$$

where f_n denotes the n-th frequency in an array consisting of N distinct frequencies. It is important to note that, in the above expression, the amplitude of each sinusoidal component is set to unity. For the sake of simplicity, all signals in this analysis are constructed from sinusoidal waves of unitary amplitude.

The parameters that are allowed to vary, and which must be considered when interpreting the results, are as follows:

- **Sampling rate** [Hz]: This parameter determines both the range of allowable frequencies that can compose the signal, as it must satisfy Nyquist's theorem, and the number of discrete data points sampled in the signal.
- **Duration** [s]: The temporal extent of the signal. Together with the sampling rate, this defines the total number of sampled points.
- **Number of components** (N): The total number of distinct sinusoidal components included in the signal.
- **Frequencies**: The frequency of each sinusoidal component, which may vary from 1 Hz up to the Nyquist frequency.

In addition to the selection of signals, the simulation of noise is also a critical aspect of the study. Two types of noise are considered:

1. **White Gaussian Noise**: This type of noise is described in Section 1.1. In Python, white noise is simulated using the *numpy.random.randn* function, which generates a sample of normally distributed random numbers [4]. To provide a controllable parameter that effectively quantifies the amount of noise added to the signal, the Signal-to-Noise Ratio (SNR) is chosen. The SNR is defined as:

$$\mathrm{SNR} = \frac{\text{signal power}}{\text{noise power}}.$$

   By providing the signal (to compute its power) and the desired SNR as input, the appropriate noise level can be computed by adjusting the noise power and scaling the Gaussian distribution obtained with *numpy.random.randn*.

2. **linear Noise**: This is defined as white noise that is modified in the frequency domain such that it varies linearly with frequency. The white noise is generated as described above, and then the Fourier Transform is computed using *numpy.fft.fft*. In the frequency domain, the noise is modified according to the following relation:

$$\text{linear\_noise\_spectrum} = \text{white\_noise\_spectrum} + \text{slope} \cdot \text{frequencies},$$

   where the slope is user-defined and the frequencies are those computed using the numpy function *numpy.fft.fftfreq*. Finally, the inverse Fourier Transform is applied, and the resulting linear noise is added to the signal.

Some considerations must be taken into account. White noise is added directly in the time domain, whereas linear noise, due to its definition, involves performing manipulations in the frequency domain. Since the signal is finite, edge effects and spectral leakage may occur. Consequently, the application of linear noise might introduce additional artifacts into the signal.

Furthermore, it is important to note that the linear noise also takes an input SNR value, which is used to compute the white noise power. However, the operation of inclining the spectrum results in a frequency-dependent power distribution that differs from the original white noise. Consequently, the resulting SNR of the linear noise may differ from the initial input SNR.

The following figures (1 and 2) present an example of the generated signal, both in its clean form and with the addition of noise, in both the time and frequency domains.

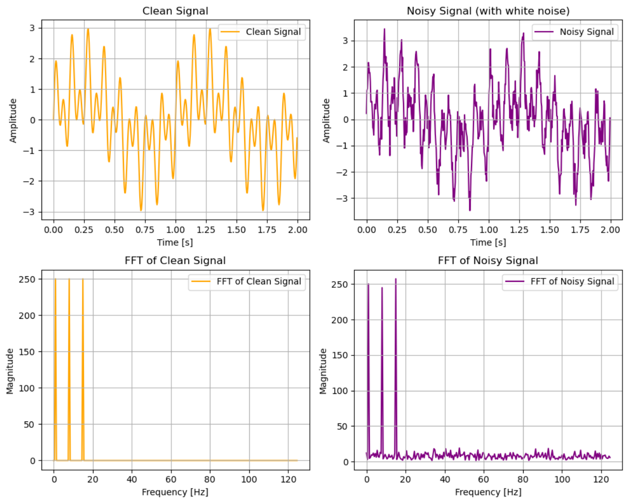


Figure 1: Example of a simulated signal affected by white noise. The left panels display the clean simulated signal, while the right panels illustrate the effect of introducing white noise.

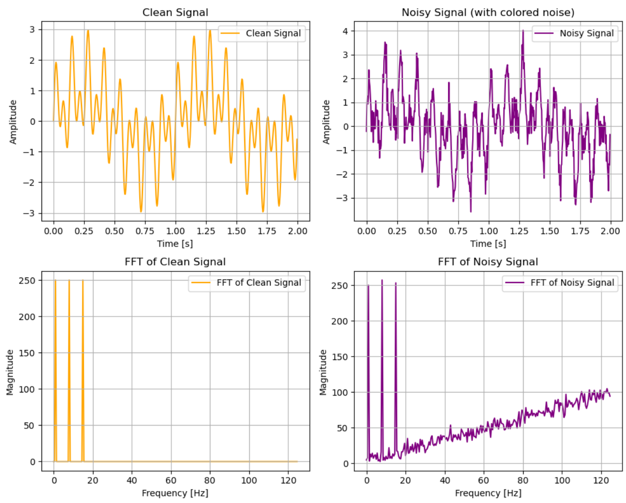


Figure 2: Example of a simulated signal affected by linear noise. The left panels display the clean simulated signal, whereas the right panels illustrate the impact of linear noise. As observed, linear noise exhibits a distinct behavior in the frequency domain, whereas its effect in the time domain is less straightforward to interpret.

In addition, the same function was implemented to generate a sum of cosinusoidal signals, following the same logic and parameters described above. In the subsequent sections, unless stated otherwise, the signal used for analyzing each approach consists of a sum of sinusoidal waves. The use of cosinusoidal signals solely introduced a phase difference; however, the function was nonetheless implemented, as certain aspects of the analysis required assessing whether a given approach was excessively influenced by the sinusoidal nature of the input data (see for example section 4.1).

Finally, a third function was implemented to generate signals composed of both sine and cosine waves. The input parameters remain the same as previously described; however, for each frequency component, it is randomly determined whether the contribution is in the form of a sine or a cosine wave.

3 Data preparation

After establishing the methodology for constructing the dataset to address the problem at hand, the next step was to identify the properties of the signals that should be exploited for the classification task described previously.

Since the definition of "same signal" is based on the frequency content, it was logical to approach the problem directly in the frequency domain. The objective was to identify features that are common among signals originating from the same process but differ for signals originating from distinct processes. Given that the noise introduced into the signals is white noise, its distinctive properties in both the time and frequency domains can be exploited.

To perform the analysis in the frequency domain, the Fourier transform was employed as the primary tool. The central idea behind this choice was to utilize the linearity of the Fourier transform. Specifically, if two signals are identical but affected by additive noise, their difference in the frequency domain should correspond to the Fourier transform of the noise. Since the noise is assumed to be white noise, the difference between the signals should ideally exhibit the characteristics of white noise itself, both in the time and frequency domains.

Mathematically, let two signals be defined as $\mathbf{X}(t) = \mathbf{x}(t) + \mathbf{n}_1(t)$ and $\mathbf{Y}(t) = \mathbf{x}(t) + \mathbf{n}_2(t)$ where $\mathbf{x}(t)$ represents the common underlying signal and $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ are the noise components affecting each signal. The difference between the signals in the frequency domain can then be expressed as:

$$\Delta\hat{X}(f) = \mathcal{F}[\mathbf{X}(t) - \mathbf{Y}(t)] = \mathcal{F}[\mathbf{n}_1(t) - \mathbf{n}_2(t)] = \hat{N}(f)$$

where $\mathcal{F}$ denotes the Fourier transform and $\hat{N}(f)$ represents the Fourier transform of the difference between the noise components. This difference should ideally retain the characteristics of white noise, as both $\mathbf{n}_1(t)$ and $\mathbf{n}_2(t)$ are assumed to be white noise processes.

Thus, by calculating the difference in the frequency domain between the two signals, it becomes possible to isolate the noise components. These components can then be analyzed to determine whether the signals are the same (in terms of the underlying signal $\mathbf{x}(t)$) or if they originate from different processes.

Another important step in the signal processing pipeline was the detrending of the transformed signal. This step was essential not only in the context of the present simulation to mitigate the effects of linear noise, as described above, but also to account for any potential frequency drift that may have been present in the data. The detrending process ensured that only the relevant signal characteristics are retained for comparison, thereby improving the reliability of the subsequent analysis.

The pipeline followed for processing the signals was as follows:

1. Compute the complex Fourier transform of the signal. To reduce redundancy, only the positive frequency components of the transformed signal are retained for further processing, as the negative frequencies carry the same information due to the symmetry of the Fourier transform. This simplifies the analysis without losing essential information.

Having computed the FFT, each of the following steps was carried out in the frequency domain.

2. Apply a linear detrending procedure to both the real and imaginary components of the FFT of each signal. This step removes any linear trends or drifts in the signal, which could obscure the underlying frequency content.

3. Compute the difference between each possible pair of transformed signals. In this step, it is crucial to track whether the two signals that originated the difference are from the same underlying process or from different processes.

4. Normalize the difference by dividing it by its standard deviation.

Due to the linearity of the Fourier transform, the resulting output was a complex-valued signal that corresponds to:

- The Fourier transform of white noise, if the original signals are identical.

- The Fourier transform of a generic signal, if the original signals differ.

A complementary procedure was also tested, where the difference between the signals was first computed in the time domain, and then the Fourier transform was applied. Although this approach is valid, it was discarded due to significantly longer computational times. This is because the number of FFTs required is much higher, as it involves computing the FFT for every pairwise combination of signals. The results obtained were identical to the primary method, but the computational inefficiency led to its exclusion. The Python code for this alternative approach is provided in the Supplementary Material folder in the GitHub repository.

4 Analysis

Multiple analysis were explored with the objective of identifying a statistical test capable of correctly detecting true positives with a probability of 95%. The central idea underlying the following analyses was to leverage specific properties of white noise. In particular, if two signals are identical, their difference should exhibit characteristics typical of white noise.

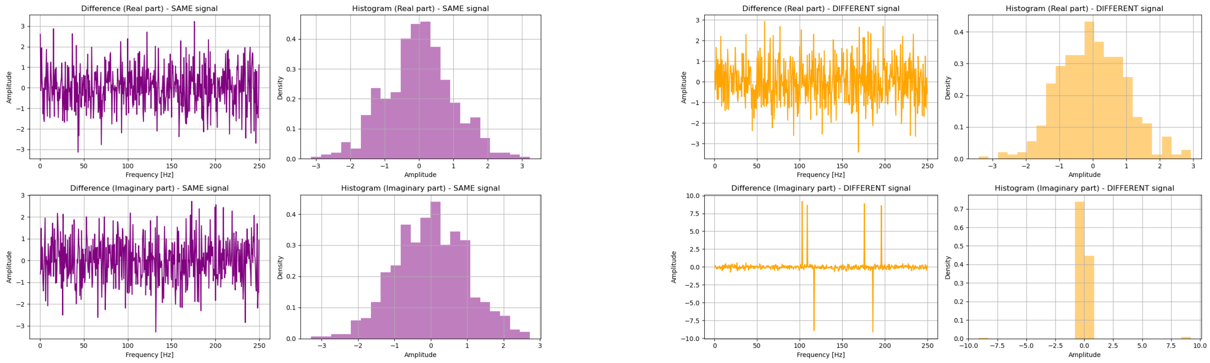
In order to ensure a valid comparison, the parameters used to generate the dataset for the statistical analysis were kept fixed. Specifically, the sampling rate was set to 500 Hz, and the signal duration was 2 seconds. A total of 20 distinct signals were generated, each with 20 noisy versions. The number of components in each signal was randomly chosen between 1 and 40 to introduce variability, and the signal-to-noise ratio was distributed between 10 and 100 dB.

The differences between signals were computed pairwise. Consequently, the sample used for the analysis presented in the following section consisted of the differences between all possible signal combinations, ensuring that repetitions were avoided.

The parameters defined above were utilized to compute the p-values and other statistical quantities, while the errors on the reported True Positive Rates were estimated using the formula $1/n$, where n represents the sample size. The analysis of signals generated with the above-mentioned parameters resulted in the p-value distributions presented in the following section (with the exception of those in section 4.3.1, where specific details are provided). Conversely, for the purpose of illustrating the amplitude distribution and presenting example plots of the differences, certain parameters were adjusted to enhance the clarity of the visual representations.

4.1 Normality of Real and Imaginary components amplitude

White noise exhibits characteristic spectral properties; in particular, the real and imaginary parts of its Fourier transform both follow a Gaussian distribution [5]. This property could have been effectively exploited, as the differences between two identical signals and two distinct signals exhibited notable discrepancies. As illustrated in Figure 6a, when two signals are identical, their difference did not present any distinct peaks and instead resembled a random process. Conversely, as shown in Figure 6b, when two signals have different frequency content, their difference revealed pronounced peaks corresponding to their constituent frequencies, along with minor variations attributable to noise.



(a) Same signal.

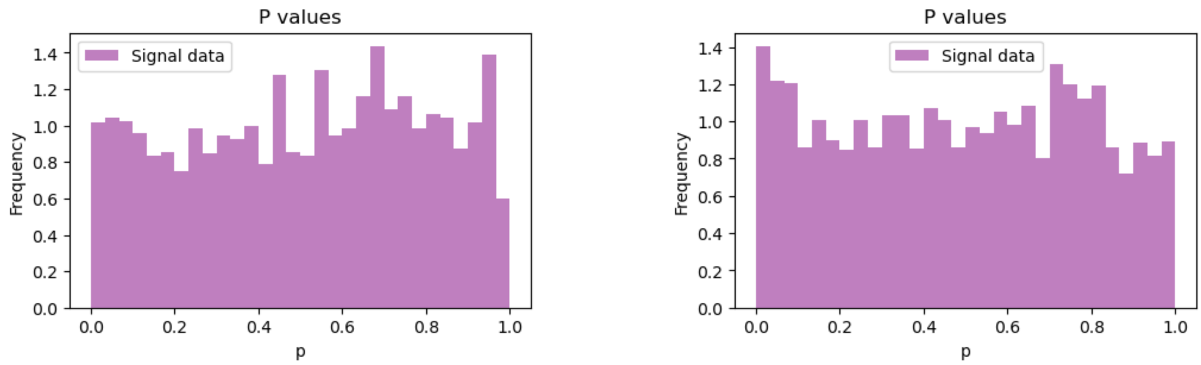
(b) Different signal.

Figure 3: Difference in the Fourier transform of two signals. In both figures, the upper panel represents the real part, while the lower panel corresponds to the imaginary part. The plots illustrate the amplitude of the difference as a function of frequency, whereas the histograms depict the amplitude distribution.

As evidenced by the plots in Figure 3, the imaginary part of the Fourier transform appeared to be more effective in distinguishing between identical and distinct signals, as the real part exhibited fewer differences between these cases. This behavior was a consequence of the specific nature of the chosen signals. As discussed in Section 2, the simulated signals are composed entirely of sinusoidal waves. Since the Fourier transform of a pure sine wave is purely imaginary, the imaginary part of the Fourier transform primarily represents sinusoidal components. To verify this observation, the same analysis was repeated using purely cosinusoidal waves, yielding the opposite result: in this case, the real part effectively captured the differences in frequency content between pairs of identical and distinct signals, as expected.

After an initial visual inspection, a statistical analysis was performed on both the real and imaginary components of the difference. To assess the normality of the data, the Shapiro-Wilk test was employed, utilizing the *SciPy* implementation of the function *scipy.stats.shapiro*. Figure 4 displays the distribution of the p values both for the imaginary and the real part.

The obtained results were consistent across both parts of the signal. In particular,



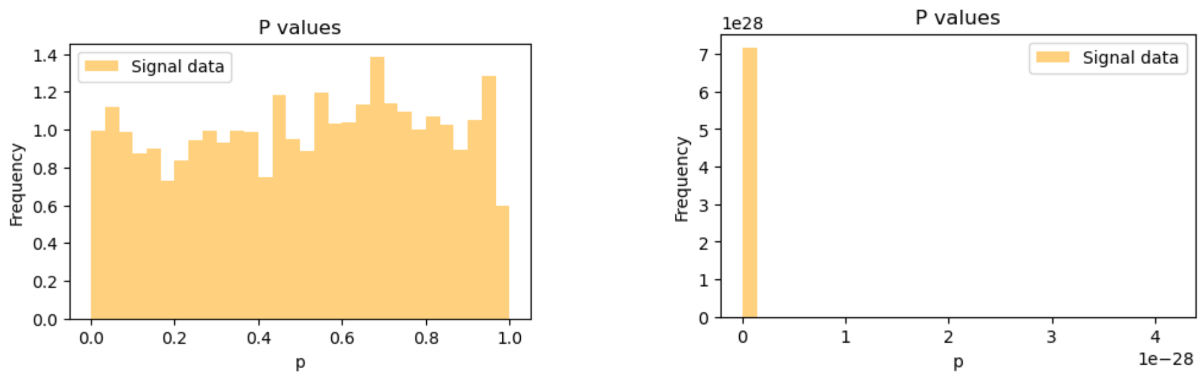
(a) P-values obtained from the Shapiro test on the real part of the difference.

(b) P-values obtained from the Shapiro test on the imaginary part of the difference.

Figure 4: Distribution of p-values derived from performing the Shapiro-Wilk test on the real (left) and imaginary (right) components of the difference of same signals.

the Shapiro-Wilk test applied to the real part yielded a True Positive Rate (TPR) of $(94.80 \pm 0.03)\%$, while for the imaginary part, the TPR is $(93.80 \pm 0.03)\%$.

Upon examining the p-value distributions for the different signals, which were expected to be mostly below the threshold of 0.05, it was observed that, for the real part, the distribution was instead approximately uniform, with p-values spread over the entire range, as shown in Figure 5. This indicates that, theoretically, 95% of the differences in the real part are classified as originating from pairs of identical signals. In contrast, the distribution for the imaginary part aligned with expectations, as anticipated. These results were consistent with the observations shown in Figure 3: since the signal was entirely composed of sinusoidal waves, the relevant frequency information is predominantly contained in the imaginary part of the spectrum [6]. The test applied to the imaginary part effectively classified two given signals, achieving a TPR of $(93.80 \pm 0.03)\%$. However, this outcome was highly dependent on the type of dataset used. For instance, when testing with cosinusoidal waves, the results were conversely opposite, with the real part of the difference correctly capturing the distinction between two signals.
Therefore, a universal recommendation for signal classification could not be drawn from these results. As a logical next step, the approach was to explore a property of white



(a) P-values obtained from the Shapiro test on the real part of the difference.

(b) P-values obtained from the Shapiro test on the imaginary part of the difference.

Figure 5: Distribution of p-values derived from performing the Shapiro-Wilk test on the real (left) and imaginary (right) components of the difference of different signals. It is important to note that the x-axis scales of the two plots differ. Specifically, the p-values shown in 5b are approximately 28 orders of magnitude smaller than those in 5a. To ensure proper visualization, a different scale was adopted.

noise that involves both the real and imaginary parts of the spectrum, thus eliminating the need to select between the two components.

4.2 Rayleigh distribution of magnitude

The Rayleigh distribution [7] is defined as the square root of the sum of the squares of two independent Gaussian components, i.e.,

$$y = \sqrt{x_1^2 + x_2^2}$$

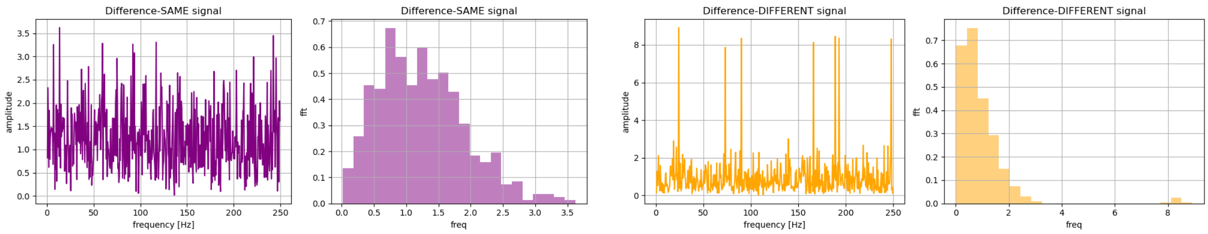
where x_1 and x_2 are independently Gaussian-distributed variables. The probability density function is given by:

$$f(x, \sigma) = \frac{x}{\sigma^2} e^{\frac{-x^2}{2\sigma^2}}$$

with σ being the scale parameter of the distribution.

Given that the objective was to identify a property of white noise that involves both the real and imaginary components of the spectrum, and knowing that these components are gaussianly distributed and independent, it followed that the magnitude of the FFT of white noise should follow a Rayleigh distribution.

The magnitude of the difference in the FFT was computed, with differences calculated pairwise as described previously, and labeled as either 'same signal' or 'different signal.' Figure 6 illustrates the amplitude and its distribution for a representative example signal. As described in the previous section, the difference between two identical signals exhibited a behavior similar to that of noise, while the difference between two distinct signals preserved the frequency content of both original signals, as evidenced by the presence of peaks.



(a) Same signal.

(b) Different signal.

Figure 6: Magnitude of the difference in the Fourier transform of two signals. The plots illustrate the magnitude as a function of frequency, whereas the histograms depict the magnitude distribution.

The statistical analysis of the distribution was conducted using the Kolmogorov-Smirnov test, implemented via the *SciPy* function *scipy.stats.kstest*. Specifically, the amplitude distribution was tested against a Rayleigh distribution, with the parameter σ estimated from the test dataset. The Kolmogorov-Smirnov test assesses whether a given set of observations follows a fully specified continuous distribution [8]. Figure 7 illustrates the resulting distribution of p-values.

As evident from the plot, the procedure resulted in an overly conservative test, as indicated by the p-value histogram exhibiting a strong bias toward 1. This suggested that the empirical distribution was systematically closer to the theoretical distribution than would be expected under a truly uniform null hypothesis.

This effect may arise from the fact that the parameters of the true Rayleigh distribution of white noise were not assumed to be known a priori but were instead estimated

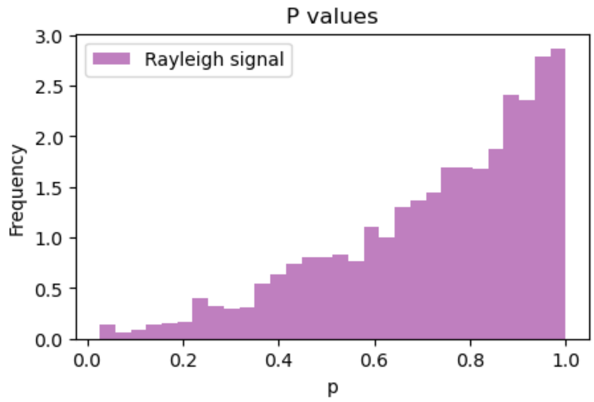


Figure 7: Distribution of p-values obtained from testing the amplitude distribution of the magnitude of the difference between the FFT of signal pairs against a Rayleigh distribution.

from the data. When certain distribution parameters are inferred from the sample, the Kolmogorov-Smirnov test tends to be conservative, meaning that the probability of a Type I error is lower than the nominal significance level reported in standard tables of the Kolmogorov-Smirnov statistic [8].

4.2.1 Focus on the most extreme value

Up to this point, the analysis focused on the entire amplitude distribution. Despite being grounded in a solid theoretical framework, the obtained results were either not generalizable (as discussed in 4.1) or excessively conservative (see 4.2).

An examination of the plots in Figure 6 revealed that the primary difference between the two cases—same signals versus different signals lay in the peak values, which characterize the frequency content and appeared as clear outliers in the histogram. The qualitative interpretation was that, for same signals, the highest values in the distribution arose from noise fluctuations and should therefore conform to a Rayleigh distribution. Conversely, for different signals, the highest amplitude values are associated with the signal itself and cannot be attributed to noise, implying that they should not follow a Rayleigh distribution.

The first approach involved testing whether the most extreme values belong to a Rayleigh distribution, specifically the one estimated from the data. In this context, the most extreme value was defined as the maximum, as the magnitude is always positive and the distinguishing characteristic between ”same” and ”different” signals is the presence of prominent peaks. To perform this test, the survival function was employed. This function provides the probability that a random variable exceeds a given value. A small p-value from the survival function suggests that the observed maximum is unlikely under the Rayleigh distribution, indicating it may be an extreme deviation. This approach is particularly useful for extreme value analysis, as it directly evaluates the likelihood of observing such large values within the assumed distribution. The plot in Figure 8 illustrates the resulting distribution of p-values.

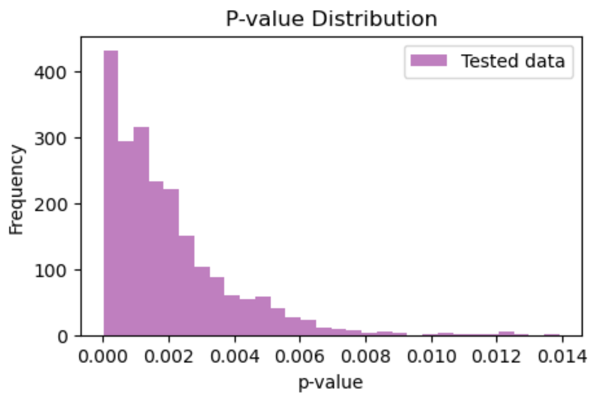


Figure 8: Distribution of p-values obtained by applying the survival function to the maximum value of the magnitude of the difference between the FFTs of pairs of identical signals.

As shown in the plot, the null hypothesis is overly rejected, indicating that either the assumption regarding the distribution the data should follow was incorrect, or that the chosen test was not suitable for the analysis. Given that the underlying theory appears sound—such as the fact that, for identical signals, both the real and imaginary components follow a Gaussian distribution, as confirmed by the Shapiro-Wilk test performed in section 4.1—it was more likely that the issue lay with the application of the test rather than the theoretical framework itself.

Since the theory behind the previous method seemed well supported, an alternative approach was explored, still relying on the idea that the maximum value of a set of identical signals should follow a Rayleigh distribution. Here, the Gumbel distribution was utilized to model the distribution of the maximum (or minimum) value from a sample of various distributions, including a Rayleigh [9]. The cumulative distribution function (CDF) of the Gumbel distribution is:

$$F(x) = exp(-exp(\alpha(x-u)))$$

where $\alpha = \frac{1.283}{\sigma(x)}$ and $u = \mu(x) - 0.45\sigma(x)$, with $\mu(x)$ representing the mean and $\sigma(x)$ the variance of the data.

As in the previous approach, the survival function was calculated using the cumulative distribution function (CDF) of a Gumbel distribution. The parameters u and α were estimated from the data, and a statistical test was conducted to assess whether the most extreme values followed the expected distribution. The resulting distribution of the p-values is shown in Figure 9.

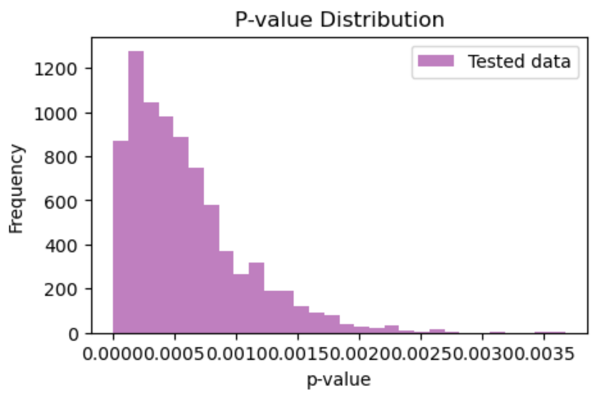


Figure 9: Distribution of p-values obtained by applying the survival function to the maximum value of the magnitude of the difference between the FFTs of pairs of identical signals, in order to test their compatibility with a Gumbel distribution.

In this case, the results consistently didn't align with expectations, as the null hypothesis is always rejected due to the small p-values. These inconsistent results may stem from an incorrect estimation of the Gumbel distribution parameters. The issue encountered with the last two methods, despite being theoretically well-supported, lay in parameter estimation. In fact, an accurate estimate of these parameters requires a high number of samples, especially since the mean is sensitive to the presence of outliers. Therefore, the parameters used to compute the survival function may not have provided a good estimate of the true parameters.

The underlying concept of this method was that, while it was acknowledged that the peak value could occasionally represent an outlier for identical signals, these instances could ideally have been confined to a 5% error rate for misclassified identical signals. However, the results indicated that relying on a single value, rather than the entire distribution, remains inappropriate, as the sample size was insufficient to perform such an analysis properly. Increasing the sample size could have been attempted, either by raising the sampling rate or extending the duration. However, the aim was to implement a test that could potentially be extended to practical applications while keeping the acquisition requirements as general as possible.

4.3 Normality of concatenated Real and Imaginary part

Among all the approaches tested, the one outlined in section 4.2 appeared to be the most promising. However, it became evident that this method was too specific for the current dataset. The optimal test was focused on assessing the normality of the imaginary part of the difference, whereas for a cosinusoidal dataset, the appropriate test should involve the real part of the spectrum. Attempting to combine both parts using the magnitude resulted in an overly conservative test. Nonetheless, an approach that incorporates both the real and imaginary components should, in principle, offer greater generalizability.

Since both the real and imaginary components followed a Gaussian distribution (and were normalized), their combined distribution should also approximate a Gaussian distribution. Therefore, the present procedure consisted of concatenating the two *numpy* arrays containing the real and imaginary parts, and then applying a Shapiro-Wilk test, as described in section 4.1, to the concatenated data. Figure 10 illustrates the distribution of p-values obtained from this procedure. The proposed approach exhibited a true positive rate of $(95.04 \pm 0.03)\%$ and produced a flat distribution, yielding a reasonable outcome.

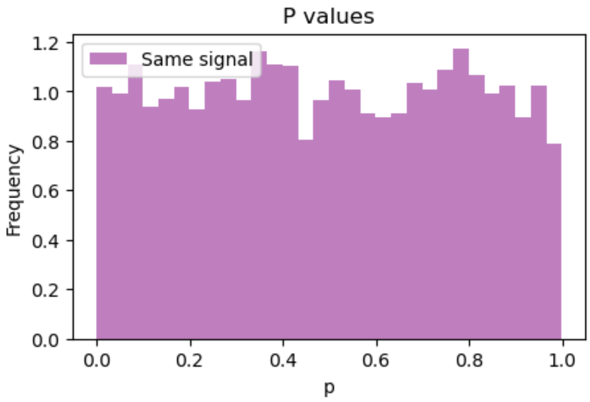


Figure 10: Distribution of p-values obtained by testing the normality of the concatenated amplitude distributions of the real and imaginary parts of the difference between the FFTs of pairs of identical signals.

To further substantiate the results, Figure 11 presents the distribution of p-values for different signals, providing additional validation for this method:

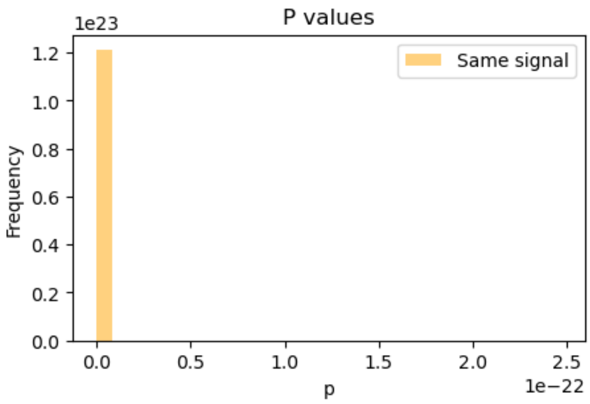


Figure 11: Distribution of p-values obtained by testing the normality of the concatenated amplitude distributions of the real and imaginary parts of the difference between the FFTs of pairs of different signals. It is important to emphasize the order of magnitude of the x-axis scale, which highlights that the p-values are consistently smaller than the significance threshold of 0.05.

4.3.1 Approach Validation

To further evaluate the validity and generalizability of the proposed approach, various signal scenarios were analyzed. All figures presented in this section depict the distributions of p-values obtained using the method described in Section 4.3. Specifically, the procedure involved computing the difference between the FFTs of pairs of identical signals, concatenating their real and imaginary components after having detrended them, and subsequently testing the resulting array for normality. This section focuses on different input signals, while the analysis methodology remains unchanged.

The following type of signals were generated:

- Sinusoidal signals, as described in Section 2, perturbed by linear noise with a spectral slope of 0.6. The corresponding results are presented in Figure 12a, yielding a true positive rate of $94.80 \pm 0.03\%$.

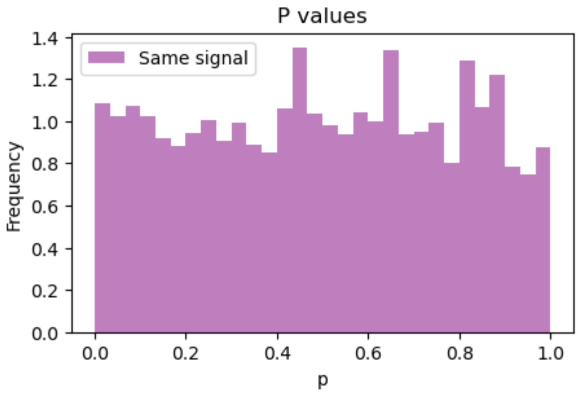
- Cosinusoidal signals affected by white noise. The dataset was constructed following the methodology outlined in Section 2, maintaining the same parameters used for sinusoidal signals. The results, shown in Figure 12b, indicate a true positive rate of $(95.00 \pm 0.03)\%$.

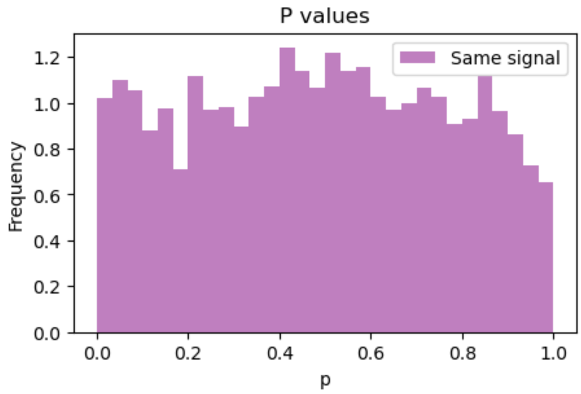
- A composite signal consisting of both sinusoidal and cosinusoidal components, affected by white noise. The results, displayed in Figure 13a, show a true positive rate of $(94.20 \pm 0.03\%)$.

- A signal composed of both sinusoidal and cosinusoidal components, perturbed by linear noise. The results are reported in Figure 13b, with a corresponding true positive rate of $(94.40 \pm 0.03\%)$

Finally, the sum of sinusoidal signals was analyzed under different parameter configurations. All parameters, except for the one under investigation, were maintained at the values specified in Section 4.

- The same parameters were used, but with a lower signal-to-noise ratio (SNR), ranging from 1 to 10, corresponding to a reduction by one order of magnitude. The

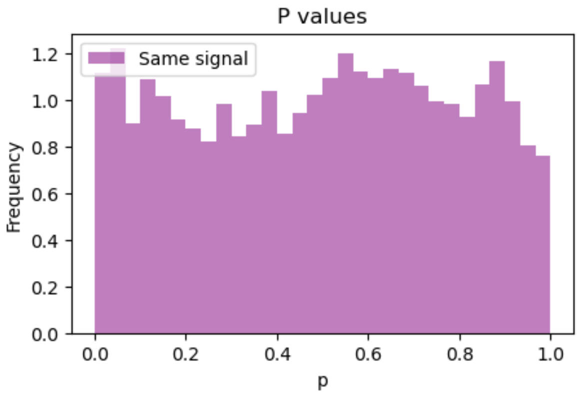


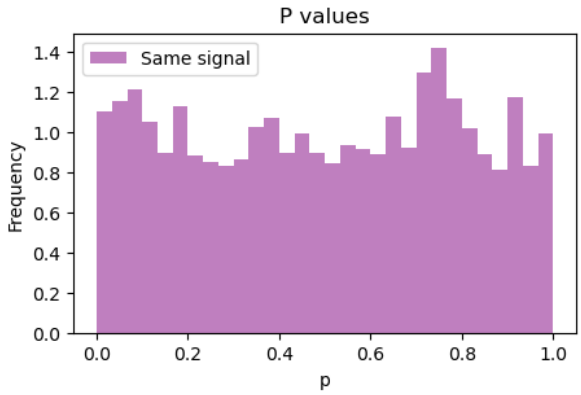


(a) Original signals are composed by a sum of sinusoidal waves affected by linear noise.

(b) Original signals are composed by a sum of cosinusoidal waves affected by white noise.

Figure 12: Resulting p value distribution applying the approach in section 4.3 to various signals type.





(a) Original signals are composed by a sum of sinusoidal and cosinusoidal waves affected by white noise.

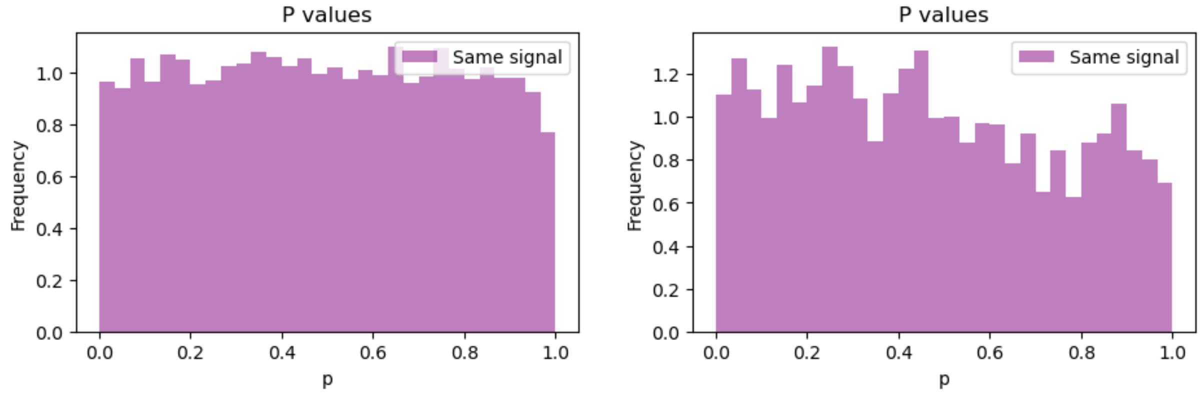
(b) Original signals are composed by a sum of sinusoidal and cosinusoidal waves affected by linear noise.

Figure 13: Resulting p value distribution applying the approach in section 4.3 to various signals type.

distribution of p-values is shown in Figure 14a, with a resulting true positive rate of $(95.30 \pm 0.03)\%$.

- An increased number of components, now ranging from 1 to 100. The corresponding o values are presented in Figure 14b, with an achieved true positive rate of $(94.2 \pm$

0.03)%



(a) Original signals are composed by a sum of sinusoidal waves affected by white noise with low SNR.

(b) Original signals are composed by a sum of up to 100 sinusoidal waves affected by white noise.

Figure 14: Resulting p value distribution applying the approach in section 4.3 to various signals type.

5 Conclusion

The objective of this study was to identify a statistical test capable of determining whether two signals are identical with an ideal true positive rate of 95%. In this context, the term "identical" refers to signals that share the same frequency content but may differ due to the presence of noise.

Leveraging the linearity of the Fourier Transform, the difference between pairs of signals was analyzed under the well-supported assumption that, if the signals are identical, their difference corresponds to the noise component. Various approaches were explored to exploit the statistical properties of white noise in a manner that facilitates the classification of signal pairs.

The method proposed as a result of this study consists of the following steps:

1. Compute the (complex) Fourier Transform of the signals to be classified.

2. Apply a linear detrending procedure to both the real and imaginary components of the transformed signals.

3. Compute the normalized difference between the two spectra.

4. Concatenate the real and imaginary components of the resulting difference.

5. Perform a Shapiro-Wilk test to assess normality. If the p-value exceeds the significance threshold of 0.05, the two signals can be considered identical with a sensitivity of $(95.04 \pm 0.03)\%$.

The proposed procedure demonstrated robustness across various conditions, including low signal-to-noise ratios, high-frequency content, and different signal types. However, a key limitation of the approach is that it requires prior knowledge of the noise affecting the signals, as it relies on statistical properties specific to white noise or linear noise, which may not generalize to other noise types.

Future work could focus on extending this methodology to different noise models, potentially transforming them into an equivalent white noise representation. Additionally, the approach could be validated on a broader range of signals, including real experimental data, to further assess its applicability in practical scenarios.

References

[1] Friedrich K. Jondral. "White Gaussian Noise – Models for Engineers". In: *Frequenz* 71.5-6 (June 2017). Received March 28, 2017; published online June 8, 2017, pp. 231–236. DOI: 10.1515/freq-2017-0064.

[2] Leiden Institute of Physics. *Noise and Signal Processing.* Accessed: 18-Feb-2025. 2003. URL: https://home.physics.leidenuniv.nl/~exter/SVR/noise.pdf.

[3] University of Pennsylvania. *Week 11: Gaussian processes - White Gaussian noise.* Accessed: 18-Feb-2025. n.d. URL: https://www.seas.upenn.edu/~ese3030/homework/week_11/week_11_white_gaussian_noise.pdf.

[4] NumPy Developers. *numpy.random.randn — NumPy v1.24 Manual.* Accessed: 2025-02-18. 2025. URL: https://numpy.org/doc/stable/reference/random/generated/numpy.random.randn.html.

[5] University of Illinois. *Lecture 3: Noise.* 2020. URL: https://courses.grainger.illinois.edu/ece417/fa2020/slides/lec03.pdf.

[6] Princeton University. *Lecture 7: Signals and Systems.* 2011. URL: https://www.princeton.edu/~cuff/ele301/files/lecture7_4.pdf.

[7] Petr Beckmann. "Rayleigh Distribution and Its Generalizations". In: *Journal of Research of the National Bureau of Standards* 68D.9 (1964), pp. 927–932. URL: https://nvlpubs.nist.gov/nistpubs/jres/68D/jresv68Dn9p927_A1b.pdf.

[8] H. W. Lilliefors. "On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown". In: *Journal of the American Statistical Association* 62.318 (1967), pp. 399–402.

[9] National Institute of Standards and Technology. *1.3.6.6.16. Extreme Value Type I Distribution.* 2012. URL: https://www.itl.nist.gov/div898/handbook/eda/section3/eda366g.htm.