

NON-PARAMETRIC DENSITY ESTIMATION BASED ON KERNEL, WAVELETS, AND K-MEANS

ESTIMATIVA NÃO PARAMÉTRICA DE DENSIDADE BASEADA EM KERNEL,
WAVELETS E K-MEANS

ESTIMACIÓN NO PARAMÉTRICA DE DENSIDAD BASADA EN NÚCLEOS,
ONDÍCULAS Y K-MEDIAS

Raquel Romes Linhares¹, Regis Nunes Vargas²

DOI: 10.54899/dcs.v23i92.6014

Recibido: 01/07/2026 | Aceptado: 02/07/2026 | Publicación en línea: 16/07/2026.

ABSTRACT

This work presents a methodology for non-parametric density estimation based on the integration of the Gaussian kernel estimator, wavelet multiresolution analysis, and clustering techniques applied to wavelet coefficients. Traditional methods often struggle to effectively separate noise from relevant density structures without relying on subjective thresholding choices. To overcome this challenge, the proposed approach initially obtains a smooth estimate using the kernel estimator. Next, this preliminary estimate is decomposed into approximation and detail wavelet coefficients. Instead of using standard thresholds, the detail coefficients are clustered using the k-means algorithm. This data-driven strategy allows for the automatic identification and separation of components associated with noise from those representing relevant underlying structures. Finally, the density function is reconstructed using only the informative coefficients. The results confirm that this methodology acts as an adaptive filtering mechanism, enabling the reconstruction of functions that preserve critical local characteristics while simultaneously reducing noise. An illustrative example was included to demonstrate the effectiveness of the method.

Keywords: Non-Parametric Density Estimation. Kernel Estimator. Wavelet. K-means.

RESUMO

Este trabalho apresenta uma metodologia para estimação não paramétrica de densidades baseada na integração do estimador kernel Gaussiano, da análise multirresolução por wavelets e de técnicas de agrupamento aplicadas aos coeficientes wavelet. Métodos tradicionais frequentemente apresentam dificuldades em separar efetivamente o ruído das estruturas relevantes da densidade sem recorrer a escolhas subjetivas de limiares. Para superar esse desafio, a abordagem proposta obtém inicialmente uma estimativa suave utilizando o estimador kernel. Em seguida, essa estimativa preliminar é decomposta em coeficientes wavelet de aproximação e

¹ Ph.D. in Mathematics, Instituto de Matemática e Estatística da Universidade Federal de Uberlândia (IME – UFU), Uberlândia, Minas Gerais, Brazil. E-mail: raquelromeslinhares@hotmail.com

² Ph.D. in Electrical Engineering, Faculdade de Engenharia Elétrica, Universidade Federal de Uberlândia (FEELT - UFU), Uberlândia, Minas Gerais, Brazil. E-mail: regisnv@yahoo.com.br

de detalhe. Em vez de utilizar limiares convencionais, os coeficientes de detalhe são agrupados utilizando o algoritmo k-means. Essa estratégia orientada pelos dados permite a identificação e separação automáticas dos componentes associados ao ruído daqueles que representam estruturas subjacentes relevantes. Finalmente, a função densidade é reconstruída utilizando apenas os coeficientes informativos. Os resultados confirmam que essa metodologia atua como um mecanismo adaptativo de filtragem, possibilitando a reconstrução de funções que preservam características locais importantes ao mesmo tempo em que reduzem o ruído. Um exemplo ilustrativo foi incluído para demonstrar a eficácia do método.

Palavras-chave: Estimativa da Densidade Não Paramétrica. Estimador de Kernel. Wavelet. K-means.

RESUMEN

Este trabajo presenta una metodología para la estimación no paramétrica de densidades basada en la integración del estimador kernel gaussiano, el análisis multirresolución mediante wavelets y técnicas de agrupamiento aplicadas a los coeficientes wavelet. Los métodos tradicionales a menudo presentan dificultades para separar eficazmente el ruido de las estructuras relevantes de la densidad sin recurrir a elecciones subjetivas de umbrales. Para superar este desafío, el enfoque propuesto obtiene inicialmente una estimación suave utilizando el estimador kernel. A continuación, esta estimación preliminar se descompone en coeficientes wavelet de aproximación y de detalle. En lugar de utilizar umbrales convencionales, los coeficientes de detalle se agrupan mediante el algoritmo k-means. Esta estrategia orientada por los datos permite la identificación y separación automáticas de los componentes asociados al ruido de aquellos que representan estructuras subyacentes relevantes. Finalmente, la función de densidad se reconstruye utilizando únicamente los coeficientes informativos. Los resultados confirman que esta metodología actúa como un mecanismo adaptativo de filtrado, permitiendo la reconstrucción de funciones que preservan características locales importantes al mismo tiempo que reducen el ruido. Se incluyó un ejemplo ilustrativo para demostrar la eficacia del método.

Palabras clave: Estimación de Densidad No Paramétrica. Estimador Kernel. Wavelet. K-means.



Esta obra está bajo una [Licencia Creative Commons Atribución- NoComercial 4.0 Internacional](https://creativecommons.org/licenses/by-nc/4.0/)

INTRODUCTION

Density estimation is one of the main tools in non-parametric statistics, being widely used when one does not want to assume a specific parametric form for the data distribution. Among the most widely used non-parametric methods, the kernel density estimator stands out (Rosenblatt, 1956). This estimator constructs a smooth approximation of the population density from the superposition of kernel functions centered on the sample observations. Its simplicity of

implementation and flexibility have made the kernel estimator one of the most popular approaches for estimating unknown densities.

Wavelet-based methods have received significant attention due to their ability to simultaneously represent global and local characteristics of functions, allowing them to capture discontinuities, asymmetries, and regions of high variability (Mallat, 2009, Linhares, 2022; Linhares and Vargas, 2025). The multiresolution representation provided by wavelets makes it possible to decompose a function into different scales, separating low-frequency components associated with the global structure from high-frequency components related to local details and noise. The use of wavelets for density estimation is a mature field, with methods based on *wavelet thresholding*, wavelet expansions, and estimation via coefficients (Donoho, et al.,1996).

In turn, the k-means algorithm constitutes one of the most widespread unsupervised clustering methods in the literature (MacQueen, 1967). Its objective is to partition a set of observations into homogeneous groups, minimizing the internal variability of the clusters. Due to its computational simplicity and efficiency, k-means has been widely used in pattern recognition, data mining, and signal processing problems.

However, the separation between wavelet coefficients associated with noise and coefficients representing relevant density structures remains a major challenge in filtering and reconstruction procedures.

In this work, a methodology is proposed that combines the initial smoothing provided by the Gaussian kernel estimator with the multiresolution representation of wavelets and the pattern identification capability of the k-means algorithm applied to the wavelet coefficients. The proposed approach seeks to use the automatic clustering of detail coefficients as an adaptive filtering mechanism, enabling the reduction of components associated with noise while preserving relevant local characteristics of the estimated density.

METHODOLOGY

Kernel Density Estimator

The kernel density estimator is one of the primary tools in non-parametric statistics for estimating probability density functions when one does not wish to assume a parametric form for the data distribution.

Consider an independent and identically distributed random sample $X_1, X_2, \dots, X_n$ coming from a distribution with an unknown density $f(x)$.

The kernel estimator is defined by

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (1)$$

where

$K(\cdot)$ is the kernel function and $h > 0$ represents the smoothing parameter, also known as the bandwidth.

In this work, the Gaussian kernel was used,

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}, \quad (2)$$

resulting in the estimator

$$\hat{f}_h(x) = \frac{1}{nh\sqrt{2\pi}} \sum_{i=1}^n \exp\left(-\frac{1}{2}\left(\frac{x - X_i}{h}\right)^2\right). \quad (3)$$

The parameter h controls the degree of smoothing of the estimate: small values produce estimates more sensitive to sample fluctuations, while larger values generate smoother curves.

In the proposed methodology, the kernel estimator is used to obtain a continuous initial approximation of the density underlying the data. This initial estimate serves as input for the wavelet decomposition step, making it possible to subsequently explore the spatial localization and multiresolution properties provided by the wavelets.

Wavelet

Wavelets constitute a mathematical tool capable of representing functions simultaneously in both the localization and scale domains, allowing for the identification of global and local data characteristics. Unlike classical Fourier series expansions, wavelets have local support, which

makes them particularly suitable for analyzing functions with discontinuities, peaks, or regions of high variability.

Multiresolution analysis is constructed from two fundamental functions: the scaling function, also called the father wavelet, and the wavelet function, frequently called the mother wavelet.

Furthermore, the scaling function and the mother wavelet satisfy fundamental properties of multiresolution analysis. The scaling function $\phi(x)$ has a unitary integral,

$$\int_{-\infty}^{\infty} \phi(x) dx = 1, \quad (4)$$

which justifies its role in representing the global structure of the function and in preserving low-frequency information. On the other hand, the mother wavelet $\psi(x)$ has a zero mean,

$$\int_{-\infty}^{\infty} \psi(x) dx = 0, \quad (5)$$

a characteristic that makes it particularly well-suited for detecting local variations, discontinuities, and oscillations in the analyzed function. Consequently, detail wavelet coefficients essentially measure local deviations relative to the global behavior captured by the scaling function. The scaling function, denoted by $\phi(x)$, is responsible for representing the low-frequency components of the function, describing its global behavior. The dilated and translated versions of the scaling function are given by

$$\phi_{j,k}(x) = 2^{j/2} \phi(2^j x - k), \quad (6)$$

where

j represents the resolution level and k represents the position of the function in the domain.

In turn, the mother wavelet function, denoted by $\psi(x)$, is responsible for describing the high-frequency components and the local structures of the function. The wavelet functions associated with the different resolution levels are defined by

$$\psi_{j,k}(x) = 2^{j/2} \psi(2^j x - k). \quad (7)$$

A function $f(x) \in L^2(\mathbb{R})$ can then be represented through the wavelet expansion

$$f(x) = \sum_k c_{J,k} \phi_{J,k}(x) + \sum_{j=J}^{\infty} \sum_k d_{j,k} \psi_{j,k}(x), \quad (8)$$

where

the coefficients

$$c_{J,k} = \langle f, \phi_{J,k} \rangle = \int f(x) \phi_{J,k}(x) dx \quad (9)$$

are called approximation coefficients and describe the global behavior of the function at resolution level J . On the other hand, the coefficients

$$d_{j,k} = \langle f, \psi_{j,k} \rangle = \int f(x) \psi_{j,k}(x) dx \quad (10)$$

are called detail coefficients and capture the local variations of the function at different resolution scales.

In the proposed methodology, the Discrete Wavelet Transform is applied to the initial density estimate obtained by the kernel estimator, producing a multiresolution representation of the density function.

In this work, the Daubechies D8 wavelet was used, implemented in the R computational environment, due to its ability to efficiently represent local structures and variations present in the signal, while simultaneously providing a smooth reconstruction of the estimated density. This decomposition makes it possible to separate information related to the distribution's structure from the components associated with noise, subsequently enabling the application of the k-means algorithm and the adaptive filtering of the detail coefficients.

K-means Algorithm

The k-means algorithm is one of the most widely used unsupervised clustering methods in the literature, aiming to partition a set of observations into K groups in such a way that elements belonging to the same group exhibit high similarity, while observations from different groups are as distinct as possible.

Formally, let $x_1, x_2, \dots, x_n$ be a set of observations and let $\mu_1, \mu_2, \dots, \mu_K$ be the centroids associated with the groups. The algorithm seeks to minimize the sum of squared distances between each observation and the centroid of the group to which it belongs, that is,

$$J = \sum_{k=1}^K \sum_{x_i \in G_k} \|x_i - \mu_k\|^2. \quad (11)$$

The iterative procedure consists of two main steps:

1. assignment of each observation to the nearest centroid;
2. updating of the centroids based on the average of the observations belonging to each group.

These steps are repeated until the algorithm converges or until there are no further changes in the composition of the groups.

In the proposed methodology, the algorithm is applied to the absolute values of the detail wavelet coefficients,

$$|d_{j,k}|,$$

considering $K = 2$ groups.

It is expected, in this way, to naturally obtain a separation between coefficients of small magnitude, generally associated with noise, and coefficients of greater magnitude, associated with relevant structures of the estimated density.

PROPOSED METHODOLOGY FOR NON-PARAMETRIC DENSITY ESTIMATION

The proposed methodology consists of five main steps: initial density estimation, wavelet decomposition, coefficient clustering, filtering, and density reconstruction.

Consider a random sample

$$X_1, X_2, \dots, X_n$$

coming from an unknown distribution with density function $f(x)$.

At first, a smooth density estimate is obtained using the Gaussian kernel estimator,

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right), \quad (12)$$

where

h represents the smoothing parameter and

$$K(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2} \quad (13)$$

is the Gaussian kernel. This step provides a continuous initial approximation of the density underlying the observed data.

After obtaining the kernel estimate, the Discrete Wavelet Transform is applied to the estimated function,

$$\hat{f}_h(x) \rightarrow \{c_{j,k}\} \cup \{d_{j,k}\}, \quad (14)$$

where

$$c_{j,k}$$

are the approximation coefficients associated with the global density structure, and

$$d_{j,k}$$

are the detail coefficients associated with the different resolution scales. The detail coefficients contain information regarding both the local characteristics of the density and the components associated with noise.

Instead of using traditional thresholding procedures, the k-means algorithm is applied to the absolute values of the detail coefficients,

$$|d_{j,k}|.$$

Considering the use of two groups, it is expected to naturally achieve a separation between small magnitude coefficients and large magnitude coefficients,

$$|d_{j,k}| \xrightarrow{\text{k-means}} G_1, G_2. \quad (15)$$

In general, coefficients of small magnitude tend to be associated with noise, whereas larger magnitude coefficients represent relevant density structures.

After clustering, the coefficients belonging to the group associated with noise are eliminated, while the coefficients considered relevant are preserved. The set of filtered coefficients is then defined by

$$\tilde{d}_{j,k} = \begin{cases} d_{j,k}, & d_{j,k} \in G_{\text{relevant}}, \\ 0, & d_{j,k} \in G_{\text{noise}}. \end{cases} \quad (16)$$

This procedure acts as an adaptive thresholding mechanism determined automatically by the clustering of the coefficients.

Finally, the density is reconstructed using the approximation coefficients and the filtered detail coefficients,

$$\tilde{f}(x) = \sum_k c_{J,k} \phi_{J,k}(x) + \sum_j \sum_k \tilde{d}_{j,k} \psi_{j,k}(x), \quad (17)$$

where

$$\phi_{J,k}(x)$$

represent the scaling functions and

$$\psi_{j,k}(x)$$

represent the wavelet functions associated with the chosen basis.

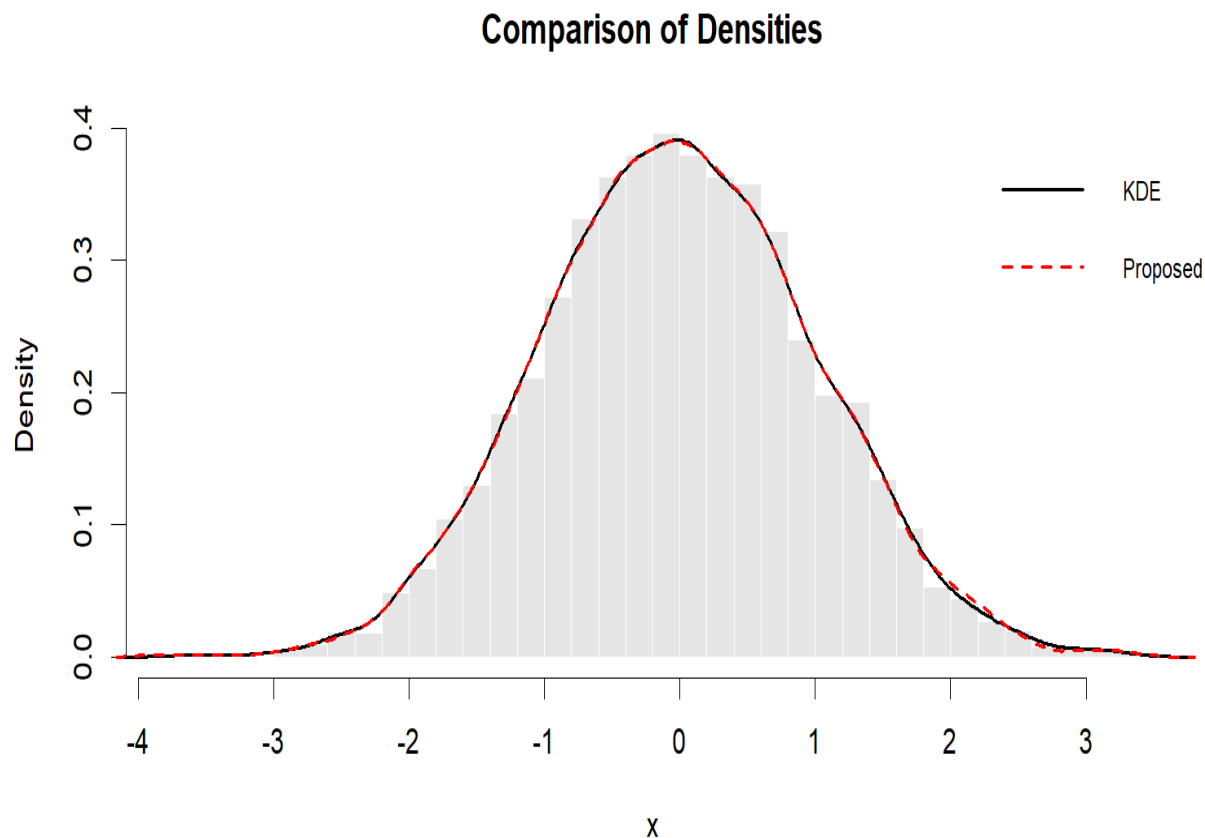
The result is a smoothed density estimate capable of preserving important local characteristics, such as peaks and regions of higher probability concentration, while simultaneously reducing components associated with noise.

APPLICATION AND GRAPHICAL ANALYSIS

The data used in this study were generated in the R computational environment, using the random seed *set.seed(22)* to guarantee the reproducibility of the results. The sample was obtained using the *rnorm(5000)* function, generating 5000 observations coming from a standard Normal distribution. It can be observed in Figure 1 that both estimates adequately reproduce the structure of the underlying density. However, the proposed methodology offers the additional advantage of performing adaptive filtering on the wavelet coefficients, allowing for the attenuation of components associated with noise and the preservation of relevant local characteristics of the distribution. In this study, the Daubechies D8 wavelet (Daubechies, 1992), implemented in the R environment, was used in the multiresolution decomposition step due to its capacity to efficiently represent local structures and variations present in the signal, while at the same time providing a smoother reconstruction of the estimated density.

Figure 1

Comparison between the density estimate from the Kernel estimator (solid black line) and the density obtained by the proposed methodology (dashed red line), using a sample of 5000 observations from a standard Normal distribution.



Source: Elaborated by the authors in R (2026).

CONCLUSION

A methodology for non-parametric density estimation was presented, based on the integration of the Gaussian kernel estimator, wavelet multiresolution analysis, and the automatic clustering of detail coefficients via the k-means algorithm. This work sought to address the challenge of separating wavelet coefficients associated with noise from those representing relevant structures of the density. The results confirm that the proposed approach effectively acts as an adaptive filtering mechanism, enabling the reconstruction of density functions that preserve critical local characteristics while simultaneously reducing noise.

From an academic perspective, this methodology contributes to the field of non-parametric statistics by introducing a data-driven alternative to traditional thresholding procedures, reducing subjective choices in the smoothing process. In terms of broader societal impact, the ability to obtain more accurate density estimations is vital for fields that rely on signal processing and pattern recognition, such as financial modeling, biomedical signal analysis, and physical data monitoring, ultimately supporting more robust decision-making processes based on empirical data.

Despite these advancements, this research has limitations. The reliance on the k-means algorithm for clustering assumes a structure that may not always be optimal for all types of density distributions, particularly those with complex multi-modal characteristics or varying noise levels. Furthermore, while the Daubechies D8 wavelet provided satisfactory results, the choice of the wavelet basis could influence the estimation quality. Therefore, recommended studies for future work include the evaluation of the methodology using different wavelet families and clustering techniques, as well as the application of this approach to real-world datasets with complex, non-standard distributions.

REFERENCES

Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM.

Donoho, D.L. et al. (1996). Density estimation by wavelet thresholding. *Annals of Statistics*.

Linhares, R.R. (2022). The ECF-WS estimator for univariate symmetric stable distributions with application in seismic trace signals. *Communications in Statistics - Simulation and Computation*, 51(12), 7487–7499.

Linhares, R.R. and Vargas, R.N. (2025). Unsupervised classification of ECG signals using DFA and wavelet entropy. *Observatório de la Economía Latinoamericana*. 23, 8 (ago. 2025), e11007.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1, 281-297. projecteuclid.org

Mallat, S. (2009). *A Wavelet Tour of Signal Processing*. Academic Press.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, 27(3), 832–837. [doi.org](https://doi.org/10.2307/2343131)