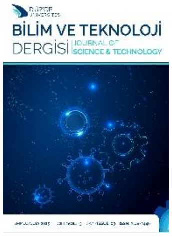

Düzce University
Journal of Science & Technology

Research Article

Removing Noise from Noisy Signal Data within Principal Component Analysis Framework

Mehmet CEVRİ[a, *]

[a] *Department of Mathematics, Faculty of Science, Istanbul University, Istanbul, TURKEY*
** Corresponding author's e-mail address: mcevri@istanbul.edu.tr*

DOI: 10.29130/dubited.1649830

ABSTRACT

The separation of noise from data represents one of the fundamental problems in signal processing. Principal component analysis (PCA) is a multivariate statistical technique that is employed in all scientific disciplines for the identification of patterns in data and the compression of data by reducing the size without significant loss of information. This paper concerns the removal of noise from noisy sinusoidal data using PCA. The aim is to achieve this by focusing on the separation of noise from signal data without estimating the parameters of sinusoidal signals. To this end, a code was developed in the Mathematica programming language, with modifications of its algorithm then being assessed on data derived from a number of noisy signals. The effectiveness of PCA was assessed by using the mean square error (MSE) values in relation to the variation in signal-to-noise ratio (SNR). The simulation results obtained demonstrate the effectiveness of PCA in removing noise from noisy sinusoidal signals.

Keywords: *Principal component analysis, sinusoidal, dimension reduction, optimization*

Temel Bileşenler Analizi Çerçevesinde Gürültülü Sinyal Verilerinden Gürültünün Giderilmesi

ÖZ

Gürültünün verilerden ayrılması, sinyal işlemenin temel problemlerinden birini temsil etmektedir. Temel bileşen analizi (PCA), verilerdeki örüntülerin tanımlanması ve önemli bilgi kaybı olmadan boyutun küçültülerek verilerin sıkıştırılması için tüm bilimsel disiplinlerde kullanılan çok değişkenli bir istatistiksel tekniktir. Bu makale, PCA kullanarak gürültülüsinüzoidal verilerden gürültünün giderilmesi ile ilgilenmektedir. Amaç, sinüzoidal sinyallerin parametrelerini tahmin etmeden sinyal verilerinden gürültünün ayrılmasına odaklanarak bunu başarmaktır. Bunun için, Mathematica programlama dilinde bir kod geliştirilmiş ve algoritmasının modifikasyonları daha sonra bir dizi gürültülü sinyalden elde edilen veriler üzerinde değerlendirilmiştir. PCA'nın etkinliği, sinyal-gürültü oranındaki (SNR) değişime bağlı olarak ortalama kare hata (MSE) değerleri kullanılarak değerlendirilmiştir. Elde edilen simülasyon sonuçları, PCA'nın gürültülü sinüzoidal sinyallerdeki gürültüyü gidermedeki etkinliğini göstermektedir.

Anahtar Kelimeler: *Temel bileşenler analizi, sinüzoidal, boyut azaltma, optimizasyon*

Received: 02/03/2025, Revised: 11/06/2025, Accepted: 12/06/2025

I. INTRODUCTION

Noisy sinusoidal signal modeling is of great importance due to its extensive applicability in numerous scientific and engineering fields. These fields encompass, but are not limited to, time series models and applications ranging from sound to radar, and nuclear magnetic resonance, and underwater acoustics [1]. The problem of removing noise from noisy sinusoids within principal component analysis (PCA) is therefore addressed here. A variety of algorithms have been employed in the literature to estimate the parameters of sinusoidal signals from noisy data and to analyze spectral data**.** The most commonly used of these are the least squares fitting [2], maximum likelihood (ML) [3], discrete Fourier transform (DFT) [4, 5], and periodogram [6]. Following the contributions of Jaynes [7], researchers in various scientific disciplines have paid considerable attention to parameter estimation within the framework of Bayesian inference. In this area, Bretthorst and colleagues [8-14] have produced seminal works.

PCA is the most frequently employed multivariate statistical technique across the full spectrum of scientific disciplines. It is a highly effective technique that can be utilized to address a multitude of issues within the fields of behavioral and social sciences, engineering [15], genetics [16, 17], neuroscience [18], and geography [19]. The advent of computing technology has facilitated the application of PCA in a multitude of fields. The initial formulation of PCA is attributed to Pearson [20], and its subsequent development is credited to Hotelling [21], who also coined the term "principal component." Currently, PCA stands as a preeminent instrument in the domains of exploratory data analysis and the construction of predictive models, as evidenced by its extensive utilization across a vast array of studies and applications, as referenced in [22-24].

Principal component analysis is a technique that can be used to clean up noisy datasets. In recent years, it has been recognized as a highly efficient tool for the analysis of high-dimensional data derived from high-spectral-resolution observational studies, as well as for the compression of redundant data [25-27]. It is evident that PCA has been applied in numerous areas of research and development, including noise filtering and data compression, as well as for the independent assessment of sensor noise [27]. The problem of noise filtering and the application of PCA to solve it are introduced in [26].

In the context of signal processing applications, PCA is typically performed on a sequence of time samples as opposed to a data set of variables. When the signal exhibits recurrent characteristics, as is the case with the automatic electrocardiogram (ECG) signal, PCA plays a pivotal role in ECG signal processing [28]. The frequency estimation of sinusoids, also known as line spectral estimation, has been the focus of research for a considerable duration. A substantial volume of research has been dedicated to this subject [29].

The present paper focuses on the question of how to detect sinusoids from noisy data in a shorter time without parameter estimation. The estimation of sinusoid parameters is known to impose a considerable computational burden and demands a significant time investment. Therefore, the impact of principal components analysis on cleaning up sinusoids from noisy data is investigated. There are many different methods for eliminating harmonic noise [30, 31]. However, certain methodologies are deemed impractical, while others encounter difficulties in the simultaneous elimination of frequency components that are in close proximity to the signal. Consequently, an approach termed random principal component analysis is proposed to address this issue [32].

The primary goal of this work was to facilitate PCA analysis of noisy sinusoids, with illustrative examples drawn from this context. It is noteworthy to emphasize, however, that the methods outlined herein find broader application in any context involving PCA analysis with noisy and/or missing data. In other words, the underlying methodology is not specific to noisy sinusoids.

II. MATERIAL AND METHOD

A. HARMONIC SIGNAL MODEL

A signal may be defined as a physical quantity that depends on time and one or more independent variables. Mathematically, a signal may be formulated as a function of one or more independent variables. In many experimental setups, the discrete data set $\mathbf{D} = \{D_1, D_2, ..., D_p\}^T$ is denoted as the output of a physical system whose model is to be developed. $\mathbf{D}$ is sampled from an unknown function $s(t;\boldsymbol{\theta})$ at discrete times $\{t_1, ..., t_p\}^T$:

$$D_i = s(t_i;\boldsymbol{\theta}) + e_i, \quad (i = 1, ..., p), \tag{1}$$

where $\boldsymbol{\theta}$ denotes a vector of unknown parameters that determine the behavior of the signal $s(t;\boldsymbol{\theta})$. The term e_i is generated from a known random process and is frequently referred to as 'noise.' When sine waves are applied to static nonlinear functions, harmonic tones are produced. The selection of the model function $s(t;\boldsymbol{\theta})$ is contingent on the application in question. In this study, the term $s(t;\boldsymbol{\theta})$ will be considered as the superposition of k sinusoids, as formulated in the following equation:

$$s(t;\boldsymbol{\theta}) = \sum_{j=1}^{k} A_j \sin(2\pi f_j t + \phi_j)\,, \tag{2}$$

where $\boldsymbol{\theta} = \left\{A_j, f_j, \phi_j\right\}$. In the context of sinusoidal signals, the variables A_j, f_j and ϕ_j are used to denote the amplitudes, frequencies, and phases, respectively, of the j^{th} sinusoidal signal.

B. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis is the most effective multivariate analysis technique for determining patterns in data and expressing them in a way that highlights similarities and differences. It is an invaluable tool for data analysis, especially in the context of large data sets that are difficult to represent graphically. PCA is a widely used technique for extracting the maximum variance from a data set, which ultimately leads to minimizing the number of variables [33, 34]. The primary objective of PCA is to determine a new set of uncorrelated variables, also termed 'principal components,' which have the capacity to explain the largest possible proportion of the total variation. PCA is a data analysis technique that can be used to extract the original signal from a set of noisy data. This process involves a series of calculations, including the singular value decomposition (SVD) and the eigenvalue decomposition of the covariance matrix. These calculations help to remove the noise from the data and reveal the underlying signal. It is employed to address the decorrelation of the signal by performing an orthogonal projection. Typically, the number of dimensions of the data is reduced from N to p ($p<N$) to eliminate undesirable components in the signal. PCA has been demonstrated to be an optimal linear dimensionality reduction technique in the mean-square sense [25]. A significant application of this technique is noise reduction, in which it is hypothesized that the data in the final components is predominantly noise. A notable strength of PCA lies in its ability to visualize multidimensional or higher-order data, where conventional methods are unable to project into a low-dimensional space, such as twoor threedimensions. PCA involves the creation of a specialized set of 'principal component' eigenvectors, which are optimized to describe the maximum variance with a minimal number of components [20, 21, 35]. For a comprehensive overview of PCA's dimensionality reduction process, refer to [36, 37].

Assume that $\mathbf{D}$ is a set of p random variables $D_1, D_2, ..., D_p$. In order to scale these variables, we subtract the mean of each dataset from each observation. This produces a dataset with a mean of zero, which allows us to identify the directions of maximum variance with greater ease. Thus,

$$d_j = D_j - \overline{D_j},\ j = 1, 2, ..., p \tag{3}$$

The random vector $\mathbf{d}^T = \left(d_1, d_2, ..., d_p\right)$ is characterized by a specific $\mathbf{\Sigma}$-covariance matrix. Consider forming new variables $Z_1, Z_2, ..., Z_k$ $(k$ □ $p)$ as a linear combination of $\mathbf{d}$-variables

$$Z_i = \boldsymbol{\alpha}_i^T \mathbf{d} = \sum_{j=1}^{p} \alpha_{ij} d_j\ ,(i = 1, 2, ..k). \tag{4}$$

PCA is a technique for dimensionality reduction from p dimensions to $k < p$ dimensions. It tries to find the most informative k linear combinations of a set of variables $Z_1, Z_2, ..., Z_k$,in sequential order.

Having defined PCs, we need to know how to find them. In the first instance, we consider the vector of random variables, $\mathbf{d}$, to have a known covariance matrix, $\mathbf{\Sigma}$. However, in a more realistic scenario, where $\mathbf{\Sigma}$ is unknown, this can be substituted for a sample covariance matrix, S. In order to ascertain the form of the PCs, it is necessary to consider first $Z_1 = \boldsymbol{\alpha}_1^T \mathbf{d}$ where the vector $\boldsymbol{\alpha}_1$ maximizes $\mathrm{Var}\left(Z_1\right) = \boldsymbol{\alpha}_1^T \mathbf{\Sigma} \boldsymbol{\alpha}_1$ subject to $\boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 = 1$. In this case, the conventional approach is to utilize the Lagrange multiplier (λ_1) technique a method frequently employed when maximizing functions subject to some constraints. To maximize $\mathrm{Var}\left(Z_1\right)$, Lagrange function $L\left(\boldsymbol{\alpha}_1, \lambda_1\right)$,

$$L\left(\boldsymbol{\alpha}_1, \lambda_1\right) = \boldsymbol{\alpha}_1^T \mathbf{\Sigma} \boldsymbol{\alpha}_1 - \lambda_1 \left(\boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 - 1\right), \tag{5}$$

where λ_1 represents a Lagrange multiplier. Upon differentiation with respect to $\boldsymbol{\alpha}_1$, the result is:

$$\left(\mathbf{\Sigma} - \lambda_1 \mathbf{I}_p\right) \boldsymbol{\alpha}_1 = 0, \tag{6}$$

where $\mathbf{I}_p$ is the $\left(p \times p\right)$ identity matrix. Hence, λ_1 is an eigenvalue of $\mathbf{\Sigma}$ and $\boldsymbol{\alpha}_1$ is the corresponding eigenvector or the weight. The objective is to maximize the given quantity:

$$\mathrm{Var}\left(\mathbf{Z}_1\right) = \lambda_1 \boldsymbol{\alpha}_1^T \boldsymbol{\alpha}_1 = \lambda_1 \tag{7}$$

Therefore, the value of λ_1 must be maximized. In this case $\boldsymbol{\alpha}_1$ represents the eigenvector associated with the largest eigenvalue of $\mathbf{\Sigma}$, whereas λ_1 denotes the largest eigenvalue itself. It can therefore be stated that $\boldsymbol{\alpha}_1$ represents the initial principal component (PC1).

In order to obtain the second PC, we want to $\mathrm{Var}\left(\mathbf{Z}_2\right)$ subject it to $\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_1 = 0$ and $\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_2 = 1$. Thus, the Lagrange function is

$$L\left(\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \lambda_2, \beta\right) = \boldsymbol{\alpha}_2^T \mathbf{\Sigma} \boldsymbol{\alpha}_2 - \lambda_2 \left(\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_2 - 1\right) - \beta \left(\boldsymbol{\alpha}_2^T \boldsymbol{\alpha}_1\right) \tag{8}$$

where λ_2 and β denote Lagrange multipliers. If the Lagrange function, as defined in Equation (6), differentiated with respect to $\boldsymbol{\alpha}_2$, it gives

$$\boldsymbol{\Sigma}\boldsymbol{\alpha}_2 - \lambda_2\boldsymbol{\alpha}_2 - \beta\boldsymbol{\alpha}_1 = 0. \tag{9}$$

If we multiply Equation (9) by $\boldsymbol{\alpha}_1^T$ from the left and use the equations $\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_2 = 0$ and $\boldsymbol{\alpha}_1^T\boldsymbol{\alpha}_1 = 1$, the value of β becomes zero. Thus Eq. (9) becomes

$$\left(\boldsymbol{\Sigma} - \lambda_2\mathbf{I}_p\right)\boldsymbol{\alpha}_2 = 0, \tag{10}$$

Hence, λ_2 once more eigenvalue of $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}_2$ the corresponding eigenvector. Again, $Var\left(\mathbf{Z}_2\right) = \lambda_2\boldsymbol{\alpha}_2^T\boldsymbol{\alpha}_2 = \lambda_2$, so λ_2 must be maximized. Similarly, the second principal component (PC2) is designated as $\boldsymbol{\alpha}_2$.

In general, the kth principal component of $\mathbf{d}$ is $\mathbf{Z}_k = \boldsymbol{\alpha}_k^T\mathbf{d}$ and $\mathrm{Var}\left(\mathbf{Z}_k\right) = \mathrm{Var}\left(\boldsymbol{\alpha}_2^T\mathbf{d}\right) = \lambda_k$, where λ_k is the kth largest eigenvalue of $\boldsymbol{\Sigma}$,and $\boldsymbol{\alpha}_k$ is the corresponding eigenvector. Therefore,the condition $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k \geq 0$ holds for the eigenvalues. Namely, the obtained principal components are in decreasing order of variance, $\mathrm{Var}\left(\mathbf{Z}_1\right) \geq \mathrm{Var}\left(\mathbf{Z}_2\right) \geq \ldots \geq \mathrm{Var}\left(\mathbf{Z}_k\right)$. In this case, $\mathbf{Z}_1$ explains as much variance as possible, and $\mathbf{Z}_2$ explains as much of the remaining variance as possible. The kth PC, $\mathbf{Z}_k = \boldsymbol{\alpha}_k^T\mathbf{d}$ maximizes $\mathrm{Var}\left(\mathbf{Z}_k\right) = \boldsymbol{\alpha}_k^T\boldsymbol{\Sigma}\boldsymbol{\alpha}_k$ subject to $\boldsymbol{\alpha}_k^T.\boldsymbol{\alpha}_k = 1$ and $\mathrm{Cov}\left(\mathbf{Z}_i, \mathbf{Z}_k\right) = 0, \left(i \neq k\right)$.

Evidence has been presented that indicates that for the third, fourth, ..., pth PCs, the vectors of coefficients $\boldsymbol{\alpha}_3, \boldsymbol{\alpha}_4, \ldots, \boldsymbol{\alpha}_k$ are the eigenvectors of $\boldsymbol{\Sigma}$ according to $\lambda_3, \lambda_4, \ldots, \lambda_p$,the third and fourth largest,...,and the smallest eigenvalue, respectively. It can thus be concluded that the eigenvalues represent the amount of variance explained by each principal component. A component with a low eigenvalue contributes only a minimal amount of variance explanation to the variables, and as such may be disregarded.

It is crucial to highlight that, on certain occasions, the vectors $\boldsymbol{\alpha}_k$ are correctly identified as *principal components* and employed to represent the directions (principal components) of the maximum variance of the data. Although this usage is occasionally defended, it is nonetheless confusing. It is therefore preferable to reserve the term 'principal *components score* **P**' for the derived variables. The principal component scores can be obtained by multiplying the centered data matrix $\mathbf{d}$ by the matrix of principal components $\boldsymbol{\alpha}$, as demonstrated in the subsequent equation:

$$\mathbf{P} = \mathbf{d}.\boldsymbol{\alpha}^T, \tag{11}$$

where $\boldsymbol{\alpha}$ refers to the *eigenvectors* and $\boldsymbol{\alpha}^T$ represents the *loadingsmatrix or principal components*. In order to ascertain the new data values, it is necessary to multiply both sides of Equation (11) by $\left(\boldsymbol{\alpha}^T\right)^{-1}$ and to consider the equation $\left(\boldsymbol{\alpha}^T\right)^{-1} = \boldsymbol{\alpha}^T$. This can be attributed to the fact that the fundamental components are all of a unit length. Moreover, it is essential to incorporate the mean value. In this case predictions for each of the original data points $\mathbf{d}_{est}$ have the following form:

$$\mathbf{d}_{est} = \mathbf{P}.\boldsymbol{\alpha} + \bar{\mathbf{d}}. \tag{12}$$

III. SIMULATION RESULTS

This section presents the findings of experimental research that demonstrates the efficacy of PCA in the removal of noise from noisy sinusoidal data. In order to comprehend the research findings, two examples are considered here. The first example involves the generation of data in accordance with the following set of guidelines:

$$d_i \;=\; \mathrm{A}\sin(2\pi \mathrm{f}\ \mathrm{t}_i + \phi) + e_i \tag{13}$$

where $A = 6,\ f = 20Hz$ and $\phi = 0\ rad$. Throughout the experiment, t_i runs in time interval 0 and 1 by 1/2000 and $e_i \sim N(0,1)$. We obtained noisy data samples ($N = 2000$) and the noisy signal stack of 20 signals is demonstrated in Figure 1. To reduce the noise, PCA is utilized to identify the principal components of the data. These components are responsible for capturing the most significant variance (patterns) while filtering out the components with less variance (which often correspond to noise).

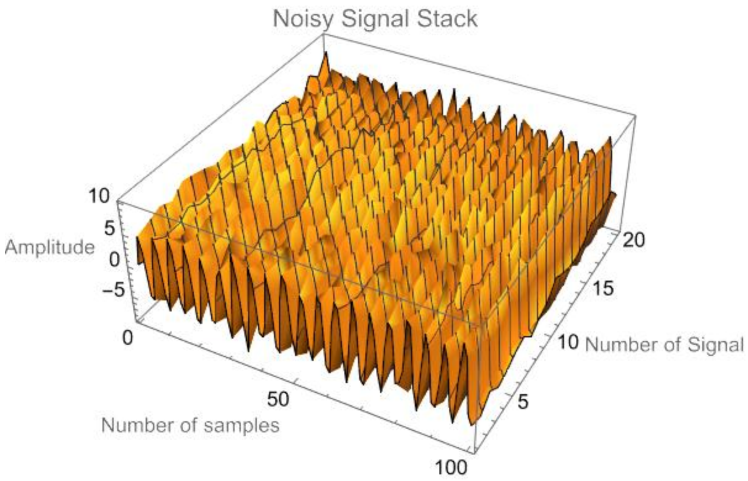

***Figure 1**. Single noisy signal stack with 20 signals*

Given the 100 samples, the data matrix is 20 x 100. All mathematical applications introduced in the paper are coded in Mathematica software. The objective is to eliminate a subset of critical components from the signal while preserving the integrity of the remaining components by performing PCA analysis. In this context, the mean value of the noise-adjusted signal is first computed. The signal is then normalized [38]. This is done by taking the mean value of each measurement in the signal and then subtracting it from the original value. Then, the covariance matrix of the signal values that have been normalized is calculated. The subsequent step in the methodology involves the determination of the eigenvalues and eigenvectors of the covariance matrix. The primary step in the denoising process entails zeroing out the basis vectors that are deemed to correspond to noise. The eigenvalue can be considered as a quantitative measure of the proportion of variation in the data that can be attributed to the principal component. A scree plot [39] can be constructed to estimate the number of eigenvalues and corresponding eigenvectors, i.e., the number of principal components, that will be utilized to obtain the denoised signal [40]. As demonstrated in Figure 2, the ordered eigenvalues are plotted against the number of components.

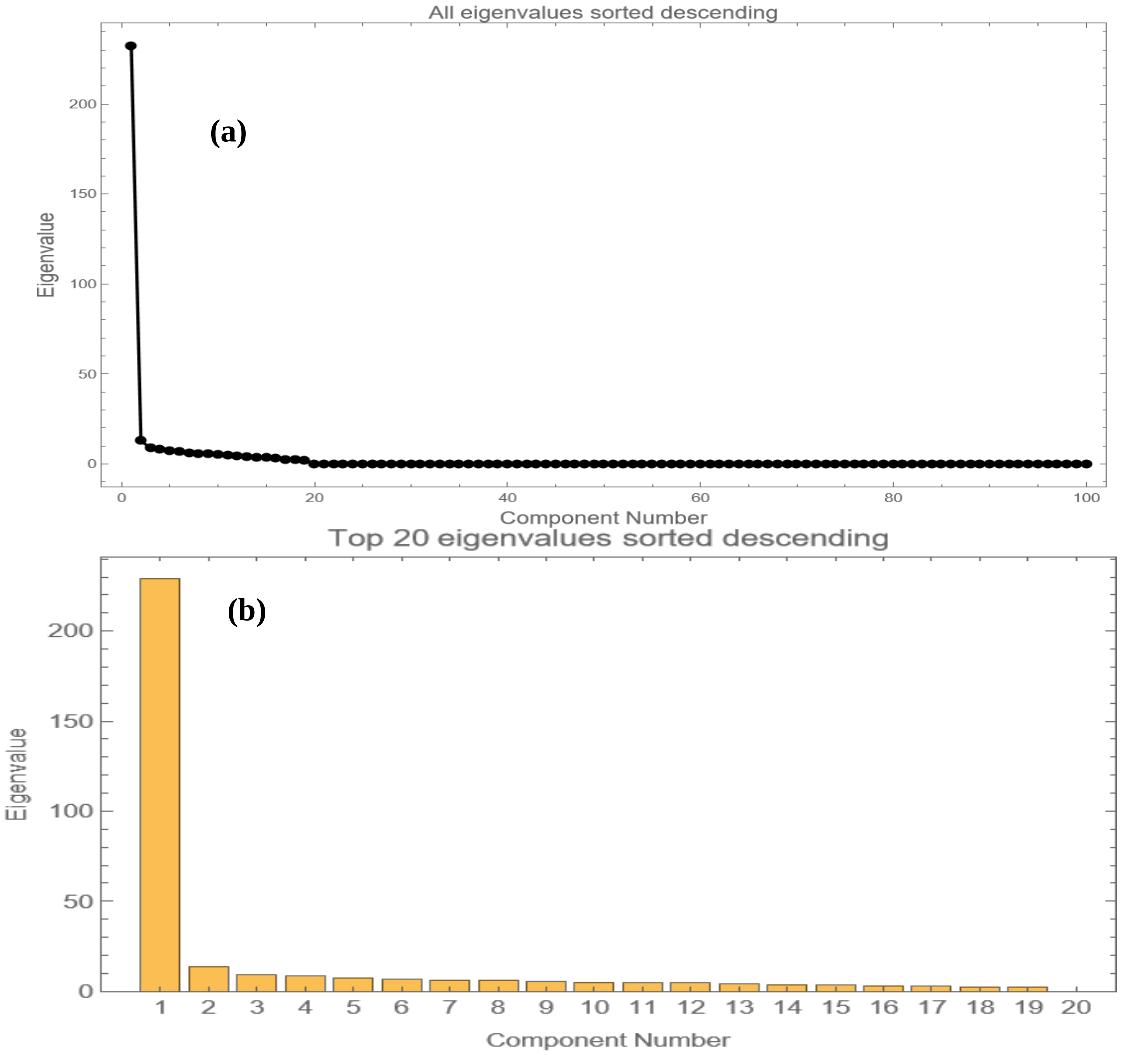

***Figure 2**. Scree plotand bar charts of the covariance matrix for data that most likely have 2 underlying components **(a)** All eigenvalues sorted descending **(b)** Top 20 eigenvalues sorted descending*

The eigenvalues presented in Figure 2 are used to detect eigenvectors, also called principal components, which in turn are evaluated to create a representation of the original signal. As demonstrated in Figure 2(a), the decreasing order of all eigenvalues is exhibited, while Figure 2(b) presents the decreasing order of the first 20 eigenvalues. As demonstrated by the figures presented, there are a number of eigenvalues that can be designated as principal components, given that their eigenvalues are considerably higher than the rest. Each component demonstrates the capacity to represent the data. As is evident from these figures, the first two eigenvalues manifest a pronounced curvature, followed by a kink and then a linear trend with a relatively shallow slope. A notable feature is the separation of the first two eigenvalues from the rest, indicating their significance in the analysis. It can thus be concluded that the eigenvectors associated with the first two eigenvalues, in order from the largest to the smallest, play a pivotal role in the noise removal process. Given the objective of removing noise from the signal and the established reliability of the first two components in representing the signal, these components were utilized to obtain the results. The remaining eigenvectors were then zeroed out, after which the modified data matrix and the sinusoidal signals were reconstructed.

PCA is a data analysis technique that identifies the components in the data (signal) that explain the biggest variance and separates them from components with lower variance (usually corresponding to noise). The reconstruction of the signal involves the following steps. Firstly, the noisy data is subjected to PCA in order to isolate the main components. Secondly, the data is then reconstructed using only the most important components (those representing the signal). The efficacy of PCA in removing noise from a signal can be illustrated by comparing the noisy signal (green) in Figure 3 and the real signal (blue) and the reconstructed signal (orange) in Figure 4. These figures demonstrate the effectiveness of PCA in smoothing out noise and restoring a signal that more closely resembles the true signal.

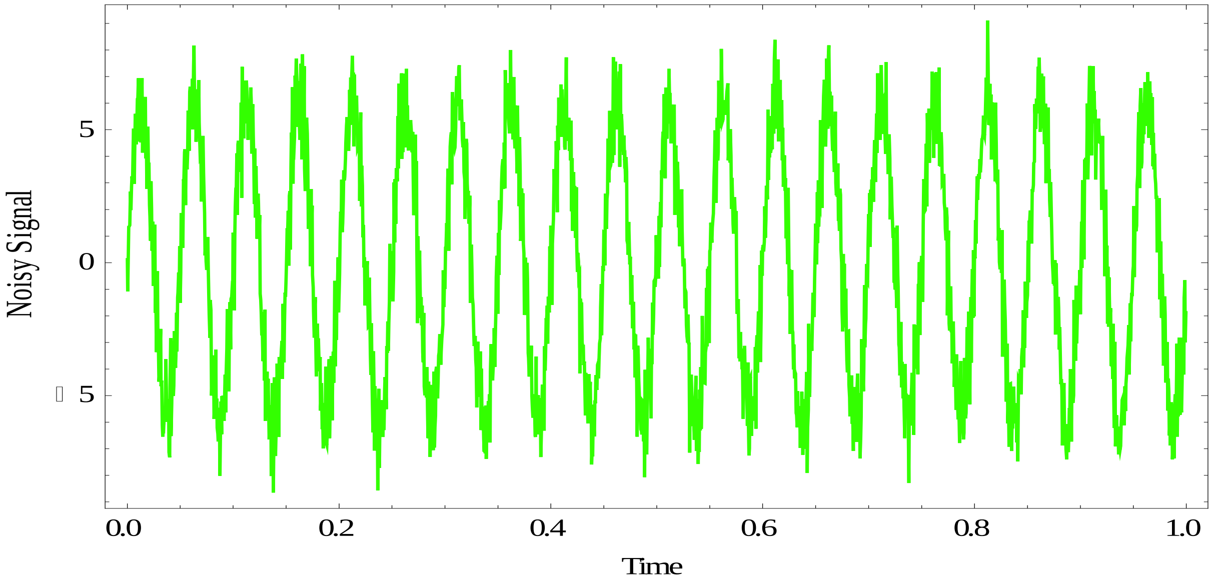

Figure 3*. Sinusoidal signal with white noise σ=1 forthe first example,*

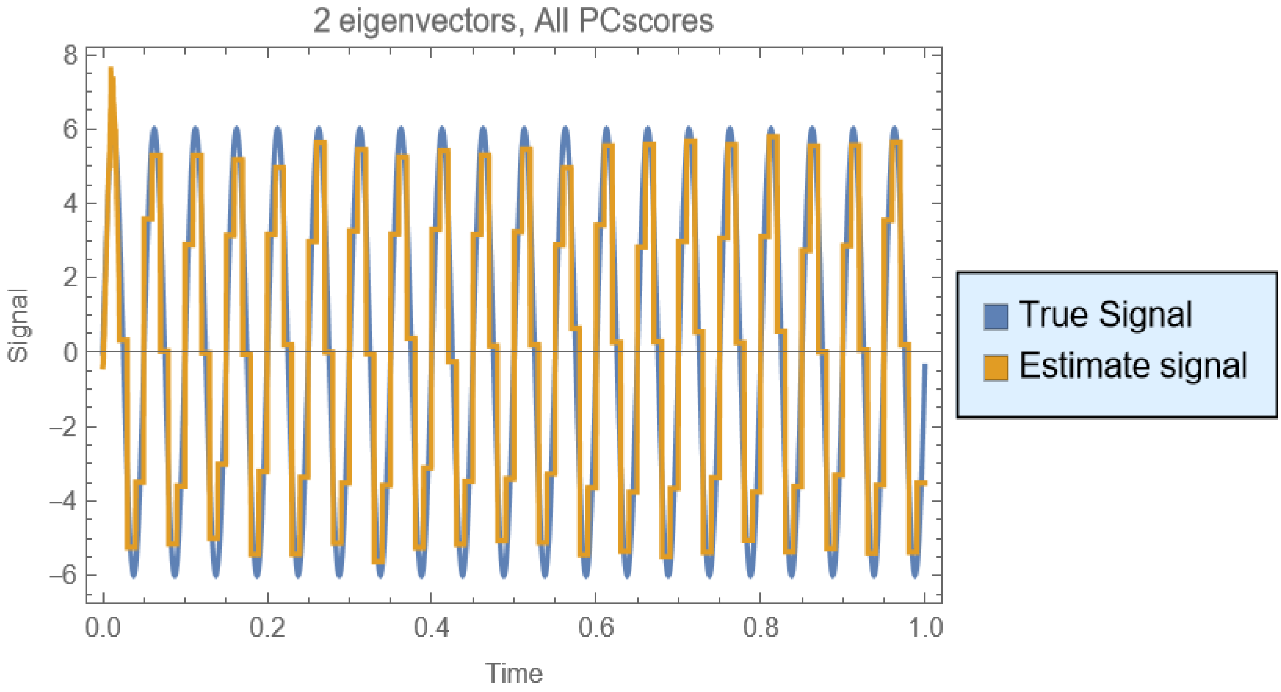

Figure 4. *The initial signal and the signal after the PCA-based noise reduction process for the first signal model*

As illustrated in Figure 4, the image obtained through PCA demonstrates a higher degree of similarity to the original image. These findings demonstrate that PCA can be utilized not only in the process of noise separation from noisy signal data but also in dimensionality reduction.

In the second experiment, multiple types of noisy data, **D**, were represented.A simple sine wave was generated with a frequency of 0.5 Hz, an amplitude of 1, and a phase of 0. Three distinct types of noise were added, generated as follows:

$$d_i = \sin(\pi \mathrm{t}_i) + e_{i1} + e_{i2} + e_{i3} \tag{14}$$

where $e_{i1} \square N(0,0.5)$, $e_{i2} \square N(0,0.3)$ and $e_{i3} \square \mathrm{P}(0.3) - 0.3$. It is evident that both e_{i1} and e_{i2} are characterized by a Gaussian distribution. These were simulated by employing random values from a normal distribution. The third noise e_{i3} is generated from a Poisson distribution. It was simulated utilizing the Poisson distribution, with subsequent adjustment to ensure centering around zero. Throughout the experiment, t_i runs in time intervals 0 and 10 by 2/100. We obtained a noisy data sample (*N=500*) and the noisy signal stack of 20 signals is shown in Figure 5.

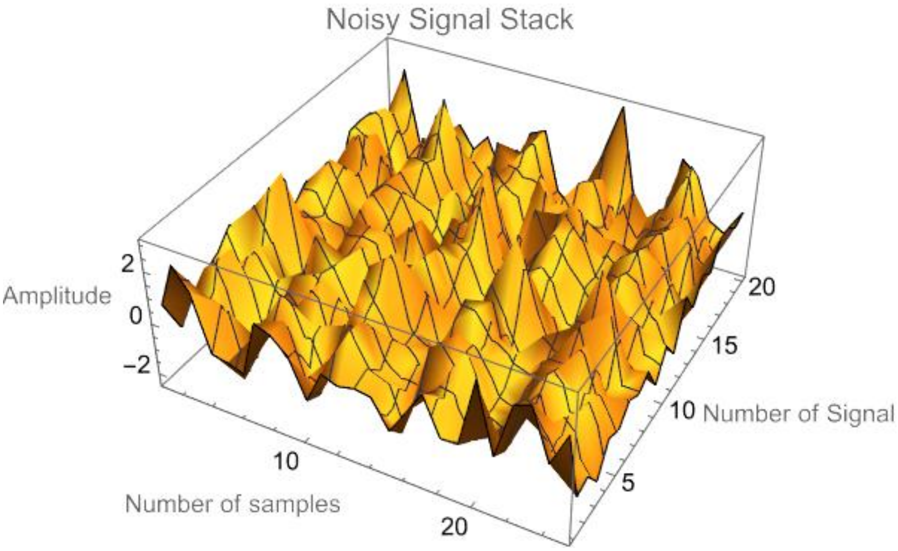

Figure 5. *Multiple noisy signal stack with 20 signals*

The ordered eigenvalues versus component number are plotted sequentially and demonstrated in Fig. 6.

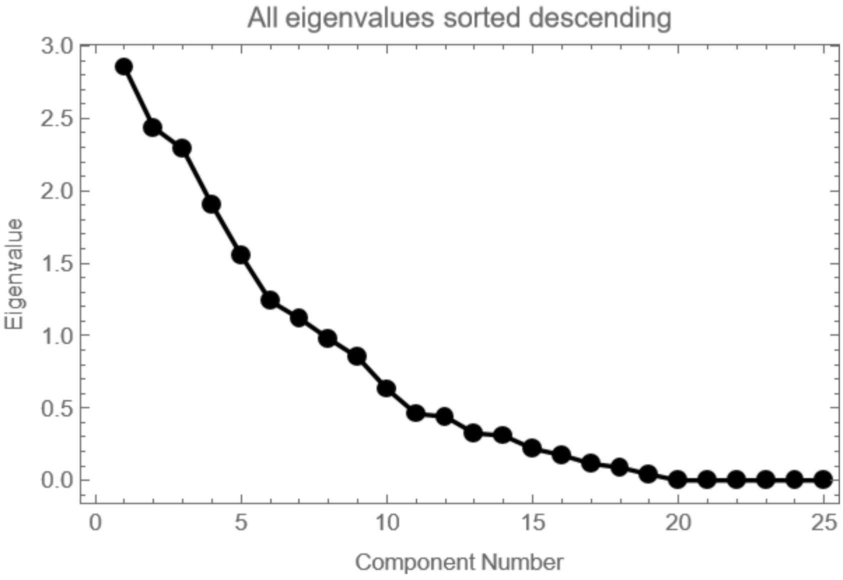

Figure 6. *Scree graph for the covariance matrix for multiple noise data*

As demonstrated in Figure 6, it is evident that the preponderance of components is deemed to be of significance in the reconstruction of the signal. Utilizing these components, the reconstructed signal is presented in Figure 7. The estimated(reconstructed) signal is then visualized alongside the noisy signal and the original(true) signal to compare the effectiveness of the noise removal. In this step, the noisy signal (in red), the original signal (in blue), and the estimated signal (in green) are plotted in Figure 7 together for comparison. The plot thus provides a visual demonstration of the effectiveness of PCA in removing noise from the original signal, thereby producing a smoother curve that more closely resembles the true signal.

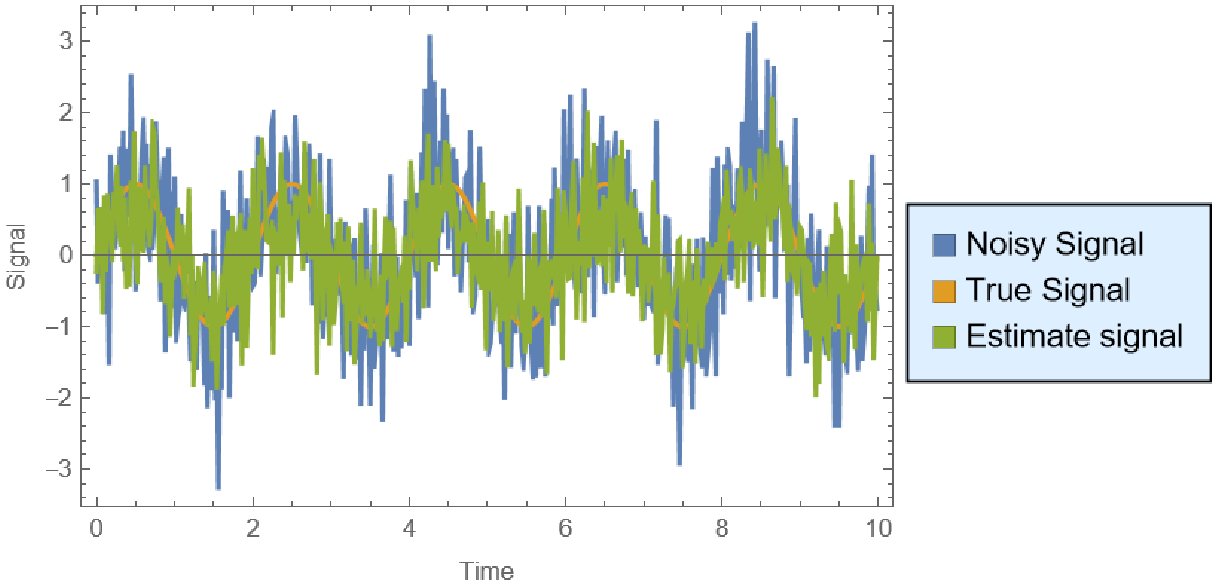

Figure 7. *Noise removal using PCA with multiple noise types*

The effectiveness of the method is assessed by calculating the mean squared error (MSE) between the true signal and the reconstructed signal. The MSE provides a quantitative measure of how well the reconstructed signal matches the true signal. A lower MSE indicates that the reconstruction is more accurate and that the noise has been effectively removed. In order to assess the impact of the de-noising method based on the PCA for sinusoidal signals, the signal- to-noise ratio (SNR) and the mean square error (MSE) must be delineated. Let α be the amplitude of signal and in this problem, we fixed α to 1 and properly scaled $e(t_i)$ to yield various SNRs, denoted as:

$$\text{SNR} = 10\log\left(\frac{\alpha^2}{2\sigma^2}\right), \tag{15}$$

and mean square error is

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(s(i) - \hat{s}(i)\right)^2, \tag{16}$$

where $s(i)$ represents the original signal, and $\hat{s}(i)$ is the de-noising signal. A total of n=500 data samples were generated from a single real-tone frequency signal model (see Eq. 14) with a variety of noise levels. MSEs of the signal were obtained. The logarithmic values were plotted in relation to the signal-to-noise ratio (SNR), which ranged from (-20) dB to (-2) dB. This is demonstrated in Figure 8. They indicate the MSE performances for different estimators. As illustrated in Figure 8, as the SNR values increase, so too do the MSE values. The minimum MSE value is 0.13, and the maximum MSE value is around 1.4, which means that the noise reduction signal has values close to the original signal.

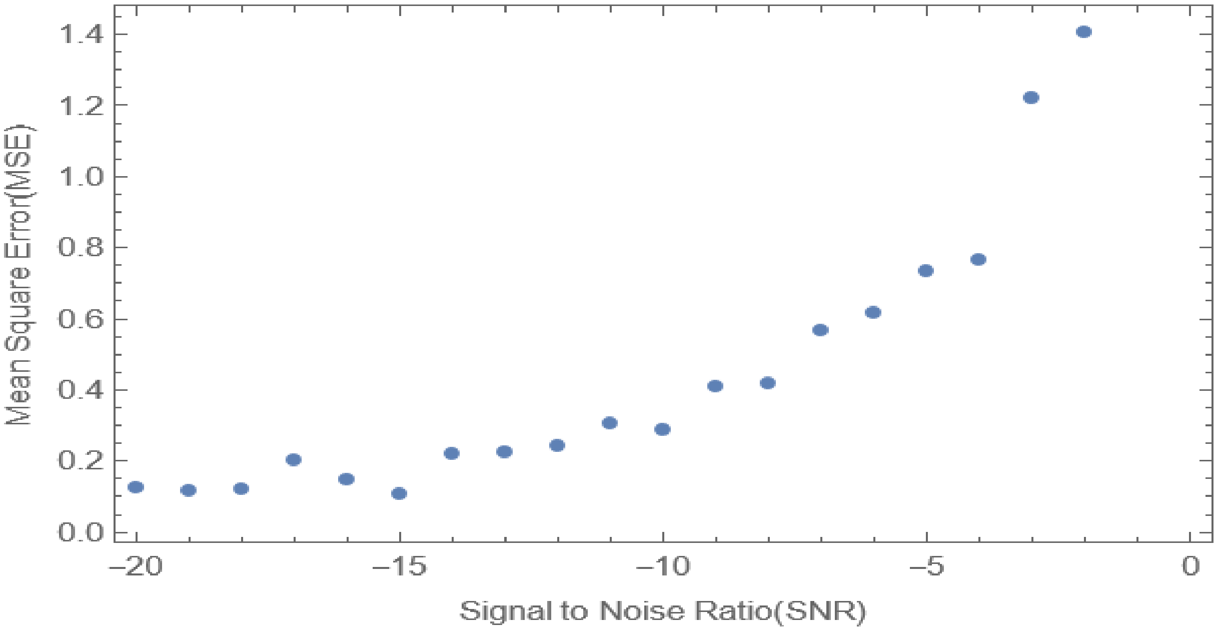

Figure 8. *MSE values with respect to SNR values for a single real- tone frequency signal model*

The MSE value for the multiple types of noisy data in Equation (14) is 0.391, indicating a slight discrepancy between the actual sine wave and the reconstructed signal. It is acknowledged that PCA-based reconstruction may not be capable of fully capturing the original sine wave; however, as the graphs presented above demonstrate, it provides a satisfactory approximation.

IV. CONCLUSION

In this paper, the PCA is utilized for the purpose of removing the noise from white, noisy, sinusoidal data. The integration of principal component analysis provides an effective strategy for reducing the dimensions of the denoised data set, resulting in the generation of a smaller number of continuous variables. The present study investigated the applicability of PCA to noisy data. The investigation focused on the ability of PCA to detect components with small variance, and this was tested on two different samples. The PCA identifies the components that explain the greatest amount of variance in the data (i.e., the signal) and separates them from the components with lower variance (often corresponding to noise).

In addition, the performance of the PCA was evaluated by examining the MSE values as a function of the change in SNR. In the instance of the SNR varying between (–20) and (–2) decibels, the MSE values are found to be relatively moderate, ranging from 0.1 to 1.4. This observation signifies the presence of a certain degree of deviation between the actual sine wave and the reconstructed signal. The results of the computer experiments demonstrate that, while PCA-based reconstruction of the sinusoidal signal may not perfectly capture the original sine wave, it does provide a reasonable approximation. A further advantage of PCA for noise removal and signal reconstruction is that it effectively cleans the data by selecting exclusively the components that capture the maximum variance (i.e., signal) and eliminating the components with minimal variance (i.e., noise). A further advantage of PCA is that it reduces the dimensionality of the data while preserving its most significant features. This process simplifies the analysis and facilitates the interpretation of the signal. Furthermore, following reconstruction, the signal approximates the true signal more closely, as a significant proportion of random noise is removed. Therefore, we may infer that PCA is employed for the purposes of noise elimination and reduction of dimensionality. While it does not inherently eliminate noise, it has the capacity to mitigate its effects.

In forthcoming studies, a comparison will be made between the performance of PCA in obtaining the original signal by separating the noise from noisy sinusoidal signals and that of other frequently used signal estimation methods in the literature.

Article Information

Acknowledgments: I would like to express their sincere thanks to the editor and the anonymous reviewers for their helpful comments and suggestions. This work is part of the project number 41895 that was submitted to Istanbul University's Scientific Projects Coordination Unit for support.

Author's Contributions: Writing—original draft, Investigation, Software, Methodology, writing—review, analysis and editing. I have read and approved the final version of it.

Artificial Intelligence Statement: No any Artificial Intelligence tool is used in this paper.

Conflict of Interest Disclosure: No potential conflict of interest was declared by authors.

Plagiarism Statement: This article was scanned by the plagiarism program.

V. REFERENCES

[1] Z. H. Michalopoulou and M. Picarelli, “Gibbs sampling for time-delay-and amplitude estimation in underwater acoustics,” *The Journal of the Acoustical Society of America*, vol. 117, pp. 799–808, 2005.

[2] R. H. Swendsen and J. S. Wang, “Replica Monte Carlo simulation of spin-glasses,” *Physical Review Letters*, vol. 57, pp. 2607-2609, 1986.

[3] R. J. Kenefic and A. H. Nuttall, “Maximum likelihood estimation of the parameters of tone using real discrete data”, *IEEE Journal of Oceanic Engineering*, vol. 12, no. 1, pp. 279–280, 1987.

[4] B. G. Quinn, “Estimating frequency by interpolation using Fourier coefficients,” *IEEE Transactions on Signal Processing*, vol. 42, no. 5, pp. 1264–1268, 1994.

[5] J. W. Cooley and J. W. Tukey, “An algorithm for the machine calculation of Complex Fourier Series,” *Mathematics of Computation*, vol. 19, pp. 297-301, 1965.

[6] N. Metropolis, A. Rosenbluth, M. Rosenbluth, A. Teller and E. Teller, “Equation of states calculations by fast computing machines,” *Journal of Chemical Physics*, vol. 21, pp. 1087-1092, 1953.

[7] E. T Jaynes, “Bayesian spectrum and chirp analysis,” in *Proceedings of the Third Workshop on Maximum Entropy and Bayesian Methods*, C. Ray Smith and D. Reidel, Eds., Boston, pp. 1-37, 1987.

[8] G. L. Bretthorst, *Lecture Notes in Statistics: Bayesian Spectrum Analysis and Parameter Estimation*, vol. 48, USA: Springer-Verlag Berlin Heidelberg, 1997.

[9] M. Cevri and D. Üstündağ, “Bayesian recovery of sinusoids from noisy data with parallel tempering,*” IET Signal Process*ing, vol 6, no. 7, pp. 673–683, 2012.

[10] L. Dou and R. J. W. Hodgson, “Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation I,” *Inverse Problem*, vol. 11, pp. 1069-1085, 1995.

[11] L. Dou, R. J. W. Hodgson, "Bayesian inference and Gibbs sampling in spectral analysis and parameter estimation II," *Inverse Problem*, vol. 11, pp. 121-137, 1995.

[12] C. Andrieu and A. Doucet, "Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC," *IEEE Transactions on Signal Processing*, vol. 47, pp. 2667-2676, 1999.

[13] D. Üstündağ and M. Cevri, "Bayesian recovery of sinusoids with simulated annealing," in *Simulated Annealing - Advances, Applications and Hybridizations*, M. S. G. Tsuzuki, Ed., Rijeka, Croatia: Intech Open, 2012. pp. 67-90.

[14] D. Üstündağ and M. Cevri, "Recovering sinusoids from noisy data using Bayesian inference withsimulated annealing," *Mathematical Computational Applications*, vol. 16, no. 2, pp. 382-391, 2011.

[15] F. A. Almeida, G. F. Gomes, P. P. Balestrassi and G. Belinato, "Principal component analysis: an overview and applications in multivariate engineering problems," *Uncertainty Modeling: Fundamental Concepts and Models, 1*th ed. A. B. Jorge, C. T. M. Anflor, G. F. Gomes and S. H. S. Carneiro, Eds., UnB, Brasilia, DF, Brazil, 2022, ch. 6, pp. 172-194.

[16] E. Elhaik, "Principal component analysis - based findings in population genetic studies are highly biased and must be reevaluated," *Scientific Reports*, vol. 12, 2022, Art. no. 14683.

[17] D. Zhang, R. Dey and S. Lee, "Fast and robust ancestry prediction using principal component analysis," *Bioinformatics*, vol. 36, pp. 3439-3446, 2020.

[18] X. Di and B.B. Biswal, "Principal component analysis reveals multiple consistent responses to naturalistic stimuli in children and adults," *Human Brain Mapping*, vol. 43, pp. 3332-3345. 2022.

[19] A. Cartone and P. Postiglione, "Principal component analysis for geographical data: The role of spatial effects in the definition of composite indicators," *Spatial Economic Analysis*, vol. 16**,** pp. 126-147, 2021.

[20] K. LIII. Pearson, "Onlines and planes of closest fit to systems of points in space," *Philosophical Magazine Series*, vol. 6, pp. 559-572, 1901.

[21] H. Hotelling, "Analysis of acomplex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 25, pp. 417-441. 1933.

[22] W. Bounoua and A. Bakdi, "Fault detection and diagnosis of nonlinear dynamical processesthrough correlation dimension and fractal analysis based dynamic kernel PCA," *Chemical Engineering Science*, vol. 229, 2021, Art. no. 116099.

[23] M. R. Mahmudi, M. R. Heydari, S. N. Qasem, A. Mosavi and S. S. Band, "Principal componentanalysis to study the relations between the spread rates of COVID-19 in high riskscountries. *Alexandria Engineering Journal*, vol. 60, no. 1, pp. 457-464, 2021.

[24] J. Song and B. ŞLi, "Nonlinear and additive principal component analysis for functional data," *Journal of Multivariate Analysis*, vol. 181, 2021, Art. no. 104675.

[25] H. Huang and P. Antonelli, "Application of principal component analysis to high-resolution infrared measurement compression and retrieval," *Journal of Applied Meteorology and Climatology*, vol. 40, no. 3, 365-388, 2001.

[26] P. Antonelli P, H. E. Revercomb and L.A. Sromovsky, "A principal component noise filter for high spectral resolution infrared measurements," *Journal of Geophysical Research*, vol. 109, 2004.

[27] D. C. Tobin, P. Antonelli, H. E. Revercomb, S. Dutcher, D. D. Turner, J. K. Taylor, R. O. Knuteson and K. Vinson, "Hyperspectral data noise characterization using principalcomponent analysis: application to atmospheric infrared sunder," *Journal of Applied Remote Sensing*, vol. 3, no. 1, 2007, Art. no. 013515.

[28] F. Castells, P. Laguna and L. Sörnmo, A. Bollmann, and J. M. Roig, "Principal component analysis in ECG signal processing," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, pp. 1-21, 2007.

[29] O. Kükrer and E. A. İnce, "Frequency estimation of multiple complex sinusoids using noise suppressing predictive FIR filter," *Digital Signal Processing*, vol. 143, 2023, Art. no. 104235.

[30] H. Karslı and D. Dondurur, "A mean-based filter to remove power line harmonic noise from seismic reflection data," *Journal of Applied Geophysics*, vol. 15, pp. 90-99, 2018.

[31] E. Shoshitaishvili, L. S. Sorenson and R. A. Johnson, "Data improvement by subtraction of high-amplitude harmonics from the 2D land vertical and multi-component seismic data acquired over the Cheyenne Belt in SE Wyoming," in *71st Annual International Meeting, SEG, Expanded Abstracts*, San Antonio, Texas, USA, 2001, pp. 2021-2023.

[32] H. Wang, H.Zhang and Y. Chen, "Sinusoidal seismic noise suppression using randomized principal component analysis, "*IEEE Access*,vol.8, pp.152131-152144, 2020.

[33] B.G. Tabachnick and L.S, Fidell, *Using Multivariate Statistics*, 4th ed., Needham Heights., MA: Pearson, USA, 2001.

[34] J. E. Jackson, *A User's Guide to Principal Components*, New York, USA: John Wiley & Sons, 1991.

[35] I. T. Jolliffe, *Principal Component Analysis*, New York, USA: Springer Series in Statistics, Springer Verlag, 2002.

[36] F. Yang, S. Liu, E. Dobriban and D. P. Woodruff, "How to reduce dimension with PCA and random projections?" *IEEE Transactions on Information Theory*, vol. 67, no. 12, pp. 8154-8189, 2021.

[37] G. Li and Y. Qin, "An exploration of the application of principal component analysis in big data processing," *Applied Mathematics and Nonlinear Sciences*, vol. 9, no. 1, pp. 1-24, 2024.

[38] R. G. Brereton, "Principal components analysis: standardisation," *Journal of Chemometrics*, vol. 39, no. 1, pp. 1-4, 2025, Art. no. e3607.

[39] R. D. Ledesma, P.V. Mora and G. Macbeth, "The scree test and the number of factors: a dynamic graphics approach," *Spanish Journal of Psychology*, vol. 18, pp.1-10, 2015, Art. no. e11.

[40] W. Min, J. Kim and O. Jo, "Denoising method for wireless communication signals based on convolutional auto encoder," in *2025 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, Fukuoka, Japan, 2025, pp. 1080-1083.