13th International Scientific Conference
Science and Higher Education in Function of Sustainable Development – SED 2023

Principal Component Analysis in Processing Photoacoustic Measurement Data

Miroslava Jordović Pavlović[1], PhD; Kristina Milojević[1]BSc; Dragana Markušev[3],PhD;Dragan Markušev[3],PhD;
[1]Western Serbia Academy of Applied Studies, departmant Uzice, Serbia, miroslava.jordovic-pavlovic@vpts.edu.rs
[1] Western Serbia Academy of Applied Studies, departmant Uzice, Serbia, kristinamilojevic123@gmail.com
[2]Institute of Physics, Laboratory for Photoacoustic, Belgrade, Serbia, dragana.markushev@ipb.ac.rs
[2] Institute of Physics, Laboratory for Photoacoustic, Belgrade, Serbia, dragan.markushev@ipb.ac.rs

Abstract:*Researchers often come across the problems of storing and processing massive data sets in machine learning tasks, as it is a time-consuming process and difficulties to interpret also arises. Not every feature of the data is necessary for predictions. These redundant data can lead to bad performances or overfitting of the model. Through this article implementation of an unsupervised learning technique, Principal Component Analysis for dimensionality reduction in preprocessing phase of photoacoustic measurement data processing is presented. It helped model deal effectively with these issues to an extent and provided sufficiently accurate prediction results.*

Keywords:*principal component analysis, simulated data, photoacoustic, measurement, neural network*

1. INTRODUCTION

Massive datasets are very common in machine learning, often with high dimensionality which includes measurements on many variables. Having a large number of features can pose many problems, and the most frequent is overfitting which reduces the ability to generalize beyond the training set. Question is how many features are irrelevant due to correlations and duplications compering with other features? Contemporary Machine Learning (ML) techniques in its preprocessing phase usually involve some of widely used mathematical procedures that reduces dimensionality. PCA is one of these mathematical procedures which main characteristic is transformation of a set of features in a dataset into a smaller number of features called principal components while at the same time trying to retain as much information in the original dataset as possible.

The Principal Component Analysis (PCA) reduction in the number of attributes is obtained by removing redundant information. PCA tries to reduce redundancy by combining the original attributes into new attributes in order to decrease the covariance, and as a result the correlation between the attributes of a data set. To do this, transformation matrix operations from linear algebra are used. These operations transform the attributes in the original data set, which can have high linear correlation, to attributes that are not linearly correlated. These are called the principal components. Each principal component is a linear combination of the original attributes. The components are ranked according their variance, from the largest to the smallest. Next, a set of components is selected, one by one, starting with the component with the largest variance and following the ranking. At each selection, the variance of the data with the selected components is measured. No new components are selected once the increase in the variance is small or a predefined number of principal components has been selected [1], [2], [3].
Advantages of using PCA are [1], [2], [3]:

- •Removes correlated features. PCA removes all the features that are correlated, a phenomenon known as multi-collinearity. If the number of features is large, finding correlated features is time consuming.
- •Improves machine learning algorithm performance. With the number of features reduced with PCA, the time taken to train ML model is now significantly reduced.
- •Reduce overfitting. By removing the unnecessary features in the dataset, PCA contributes to overcome overfitting.

On the other hand, PCA has its disadvantages [1], [2], [3]:

•Independent variables are now less interpretable. Each of PCA reduced component isnow a linear combination of original features, which makes it less readable and interpretable.
•Information loss. Data loss may occur in case of not proper choice the right number of components.
•Feature scaling. Because PCA is a variance maximizing technique, PCA requires features to be scaled prior to processing.

The article presents discussion on a convenience of use of PCA technique in processing photoacoustic measurement data.
Photoacoustics is one of phototermal methods which aim is physical characteristics determination of measured sample. It is experimental life science that has, in the recent years, an explosion of the data available from experiments, especially a simulated experimental values or numerical experiments. In our previous work, [4] the potential of using simulated data in photoacoustics is proved and became practice in designing machine learning models for decision making. A mainstay of simulated data generation is developed theoretical-mathematical model of photoacoustic response. Validity of created data is obtained due to experimental experience and expert knowledge. Numerical experiments imitate proximately all experimental conditions. Hundreds of measurements for a single experiment are reported and therefore the statistical methods face challenging tasks when dealing with such high dimensional data.

2. PROBLEM DESCRIPTION

Database that we investigate in the paper consists of 270 000 records. Each record has 200 instances of amplitude and 200 instances of phase obtained by sampling amplitude and phase characteristics of simulated photo acoustic response at defined points of frequency axe [4]. Photoacoustic response hides, among other harmful influences, the influence of measurement system, represented as distortions. The main part of measurement system in the photoacoustic experiment has microphone as a detector [5]. Without going into deeper analysis and explanations of photoacoustic experiment, some general truths will be represented here. Microphone has the greatest impact to distortions, its response in the frequency and time domain differs due to construction, geometry and membrane type, and two identical microphones do not exist in practice. Having these facts in mind it is obvious why a lot of our research attention is focused to the correction of distorted photo acoustic signal where distortions are mainly caused by microphone [6] [7].

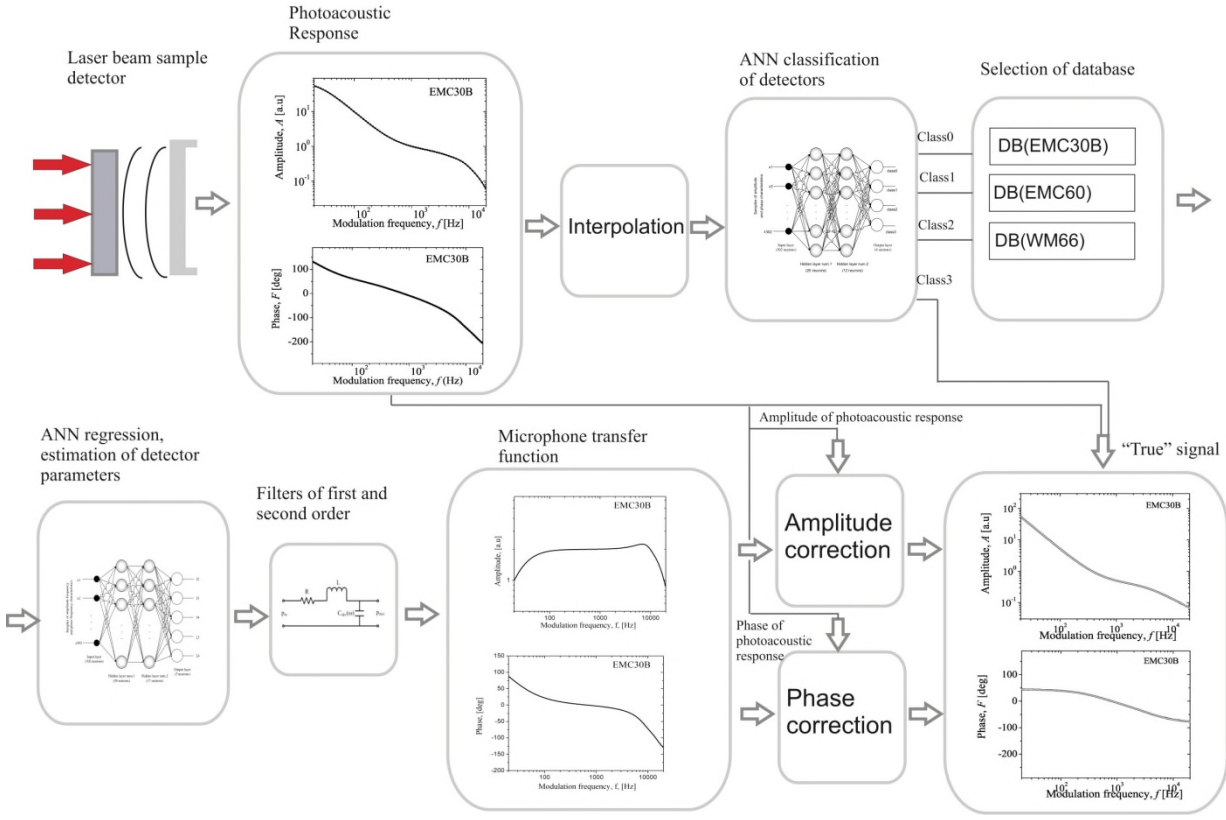

Figure 1: The block sheme of distorted photoacoustic measurement signal autocorrection method [6]

In Figure 1, a block scheme of the autocorrection of distorted photoacoustic signal is presented. The PA response of the experimental setup measured in a certain number of points represents the input signal. Input signal is first interpolated to some points that correspond to the number of neural network input vector elements. The interpolated PA response is an input of the classification model. The output is the type of microphone used as a detector in the setup. Further more, class of microphone determines the corresponding database. Regression model, trained on the selected database, gives the targeted microphone parameters. Results of corrections are amplitude and phase characteristic of the "true" PA signal. In the case of class 3 (IM) regression is not executed, experimental signal is the "true" PA signal. Detail analysis of the method is explained in our previous work [6]. In this paper focus is dimensionality reduction of the input vector.
The dataset is generated using well defined preset of three electret microphones frequently used in photoacoustic experiment and ideal microphone (microphone of ideal characteristics) used as a reference [4]. Visualization of the dataset in the form of scatter diagram, is given in Figure 2. Each point on a scatter diagram is one point of 200 points that corresponds to one curve of 270,000 curves in the database. Different classes of microphone are presented with different colors.

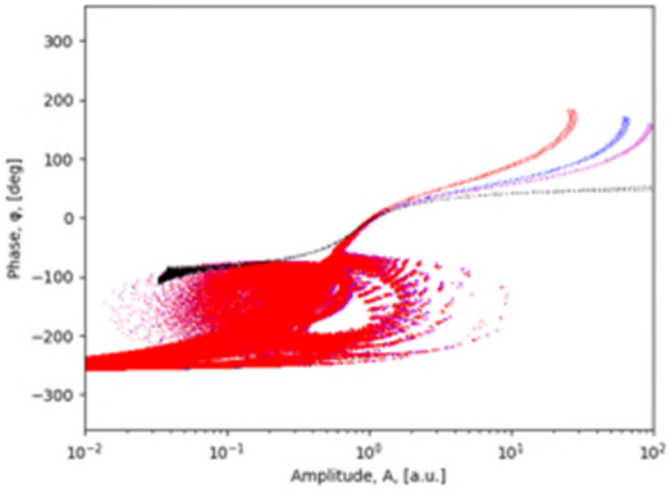

Figure 2: Scatter diagram of the data set [4]

3. RESULTS AND DISCUSSION

PCA applied on whole dataset (400 features) provides promising results of retained variance, Figure 3. It means that the dimensionality reduction could be done even to 2 components, where retained variance is 99.55%. Results for 4 and 6 components are 99.89%, 99.96% respectively.

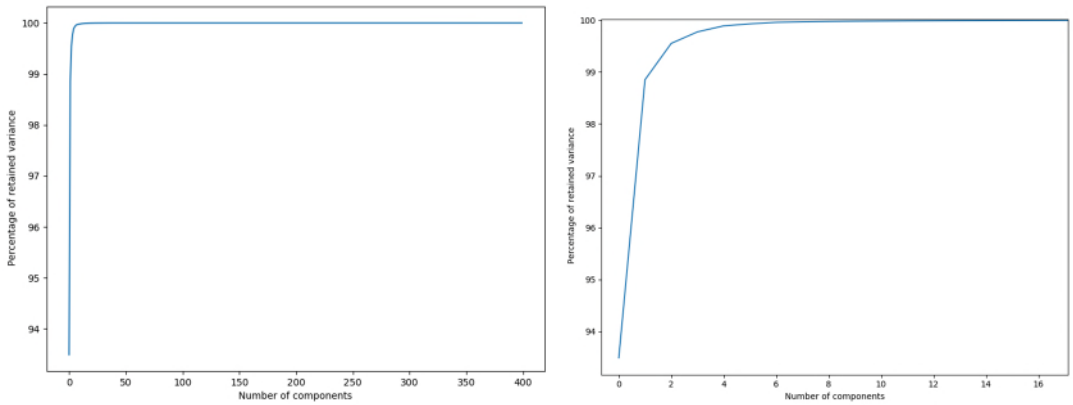

Figure 3: Retained variance in realtion to number of components

Regarding literature suggestions [8] [9] [10], 99% of retained variance means preserved information in the data set, it can be concluded that the idea of dimensionality reduction of input vector is justified. Furthermore the decision on the number of PCA components could be choice of nearest result to 99% and that means 2 components.
Validity of proposed dimensionality reduction is checked on classification model for recognition of microphone type perfomance.Input vector is reduced in accordance with proposed dimensionality reduction.

Training, validation and test sets are obtained randomly because dataset is first shuffled and then divided into training, validation and test set. Generalization of the results is obtained on that way, thus 243 000 records or 90% of the total number of records belongs to the training set, 13500 records or 5% belong to the validation set and the rest belongs to the test set.
Comparison of the classification model performance with [4] and without of dimensionality reduction done in preprocessing phase is presented in Table 1. Analyzing the table it can be concluded that the accuracy of the model is slately lower in the case of dimensionality reduction but it is still a very good result. The implementation of PCA technique is thus proven to be the adequate dimensionality reduction technique regarding the PA experiment request for precision and real time work.

	Trainaccuracy(%)	Dev accuracy(%)	Test accuracy(%)	Number of epochs	Prediction time (ms)
Perfomance of the model without dimensionality reduction	99.99	99.99	99.99	100	14ms
Perfomance of the model with dimensionality reduction	99.21	99.12	99.15	100	13ms

Table 1: Performance of the classification model with and without the dimensionality reduction

The reliability of the classification model is tested on independent data tests. Sixteen different independent data sets, meaning for different amplitude and phase characteristics for each type of microphone were created, where the microphone parameter values differed from those on which the network was trained, but in the given parameter range. Results are presented in Table 2. According to table our model is reliable, it recognizes the microphone type precisely and gives an answer regarding the microphone type in real time.

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Class.	1	3	0	2	1	2	3	0	2	3	1	2	1	0	3	0
Accuracy	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T	T

Table 2: Indepedent data tests, T means true

4. CONCLUSION

The paper discusses the results of PCA technique application in the preprocessing phase of PA measurement data processing. The change in performance of classification model is negligible compering to classification model without dimensionality reduction.It can be concluded thatPCA is suitable for implementation regarding PA experiment request for precision and real time work. Managing and exploratory analysis of the dataset are now much easier. However, interesting question came up during this research. Is there a possibility of reduction in the number of measurement points and thus simplifying the measurement procedure in real experiments? Such a reduction would significantly reduce the measurement time and researcher engagement. Unfortunately, PCA is not suitable for solving this problem. PCA component is a linear combination of original features, so the number of features is not changed, the same as number of measurement points, but its interpretation is. Solution of araised problem will need further investigation.

REFERENCES

[1] Pavel Pořízka, Jakub Klus, Erik Képeš, David Prochazka, David W. Hahn, Jozef Kaiser, „On the utilization of principal component analysis in laser-induced breakdown spectroscopy data analysis, a review“, Spectrochimica Acta Part B: Atomic Spectroscopy,Volume 148, 2018,Pages 65-82, ISSN 0584-8547, https://doi.org/10.1016/j.sab.2018.05.030.
[2] Konishi, T., Matsukuma, S., Fuji, H. *et al.* Principal Component Analysis applied directly to Sequence Matrix. *Sci Rep*9, 19297 (2019). https://doi.org/10.1038/s41598-019-55253-0
[3] Daoqiang Zhang, Zhi-Hua Zhou, (2D)2PCA: Two-directional two-dimensional PCA for efficient face representation and recognition, Neurocomputing, Volume 69, Issues 1–3, 2005, Pages 224-231, ISSN 0925-2312, https://doi.org/10.1016/j.neucom.2005.06.004.

[4] Jordović-Pavlović M., Kupusinac A., Galović S., Markushev D., Nešić M., Djordjević K., Popović M., Potential of Using Simulated Data in Processing Photoacoustic Measurement Data. In: Proceedings of 8th In- ternational Conference on Electrical, Electronic, and Computing Engineering (IcETRAN), ISBN 978-86- 7466-894-8 (2021),
[5] M. D. Rabasovic, M. G. Nikolic, M. D. Dramicanin, M. Franko, and D. D. Markushev, "Low-cost, portable photoacoustic setup for solid samples," Meas. Sci. Technol., vol. 20, no. 9, p. 95902, 2009.
[6] M. I. Jordovic-Pavlovic et al., "Computationally intelligent description of a photoacoustic detector," Opt. Quantum Electron., vol. 52, no. 5, pp. 1–14, 2020.M. Nesic et al., "Development and comparison of the techniques for solving the inverse problem in photoacoustic characterization of semiconductors," Opt. Quantum Electron., vol. 53, no. 7, p. 381, 2021.
[7] M. I. Jordović-Pavlović, M. M. Stanković, M. N. Popović, Ž. M. Ćojbašić, S. P. Galović, and D. D. Markushev, "The application of artificial neural networks in solid-state photoacoustics for the recognition of microphone response effects in the frequency domain," J. Comput. Electron., vol. 19, no. 3, pp. 1268–1280, 2020.
[8]Shalev-Shwartz, S. and S. Ben-David, *Understanding machine learning: From theory to algorithms.* Under- standing Machine Learning: From Theory to Algorithms, 2014: p. 1-397.
[9] Bishop, C.M., Pattern Recognition and Machine Learning. 2006
[10] Martel, E., et al., Implementation of the Principal Component Analysis onto high-performance computer facilities for hyperspectral dimensionality reduction: Results and comparisons. Remote Sensing, 2018.
[11] M. Nesic et al., "Development and comparison of the techniques for solving the inverse problem in photoacoustic characterization of semiconductors," Opt. Quantum Electron., vol. 53, no. 7, p. 381, 2021
[12] M. V. Nesic, M. N. Popovic, S. P. Galovic, K. Lj. Djordjevic, M. I. Jordovic-Pavlovic, V. V. Miletic, and D. D. Markushev, Estimation of linear expansion coefficient and thermal diffusivity by photoacoustic numerical self-consistent procedure, March 2022, Journal of Applied Physics, https://doi.org/10.1063/5.0075979
[13] K.Lj. Djordjević, S.P. Galović, M.N. Popović, M.V. Nešić, I.P. Stanimirović,Z.I. Stanimirović, D.D. MarkushevUse Neural Network in Photoacoustic Measurement of Thermoelastic Properties of Aluminum foil, Measurement (2022) 111537, 0263-2241https://doi.org/10.1016/j.measurement.2022.111537
[14] K. Lj. Djordjevic, S. P. Galovic, M. I. Jordovic-Pavlovic, Z. M. Cojbasic, D. D. Markushev, Improvement of Neural Networks Applied to Photoacoustic Signals of Semiconductors With Added Noise, Silicon,https://doi.org/10.1007/s12633-020-00606-y.