References:

1. Malyushevskaya, A., Koszelnik, P., Yushchishina, A., Mitryasova, O., Mats, A., & Gruca-Rokosz, R. (2023). Eco-Friendly Principles on the Extraction of Humic Acids Intensification from Biosubstrates. *Journal of Ecological Engineering,* 24(2), 317–327. https://doi.org/10.12911/22998993/156867

2. Shablia, V. P., & Tkachova, I. V. (2020). Machine and manual working actions for different manure removing technologies. *Boletim de Indústria Animal. Instituto de Zootecnia. Nova Odessa. Brasil*, 77, 1–14. https://doi.org/10.17523/bia.2020.v77.e1482

3. Xiao-xia, Guo, & Hongtao, Liu (2019). Humic substances developed during organic waste composting: Formation mechanisms, structural properties, and agronomic functions. *Science of The Total Environment*, 662. https://doi.org/10.1016/j.scitotenv.2019.01.137

UDК: 57.087.1

OVERVIEW OF CLUSTERING OF BIOLOGICAL SAMPLES USING PRINCIPAL COMPONENT ANALYSIS

Somriakov Bohdan, *a student*
Petro Mohyla Black Sea National University,
Mykolaiv, Ukraine

Clustering [1] of biological samples using Principal Component Analysis (PCA) [2] is important for simplifying complex data, revealing hidden structures, and enhancing the interpretation of biological patterns. This approach supports more informed research and decision-making in biology and medicine.

In K-means clustering [3], each data point xi is assigned to the nearest centroid ci based on the Euclidean distance [4] which is demonstrated within formula 1. Formula 2 shows how the centroids [5] are recalculated as the mean of all points assigned to each cluster.

$$d(x_i, c_j) = \sqrt{\sum\nolimits_{p=1}^{n} \left(x_{i,p} - c_{j,p}\right)^2} \qquad (1)$$

$$cj' = \frac{1}{|C_j|} \sum\nolimits_{x_i \in C_j} x_i \qquad (2)$$

Figure 1 shows an example output for K-means clustering, illustrating how data points [6] are grouped into clusters with distinct centroids.

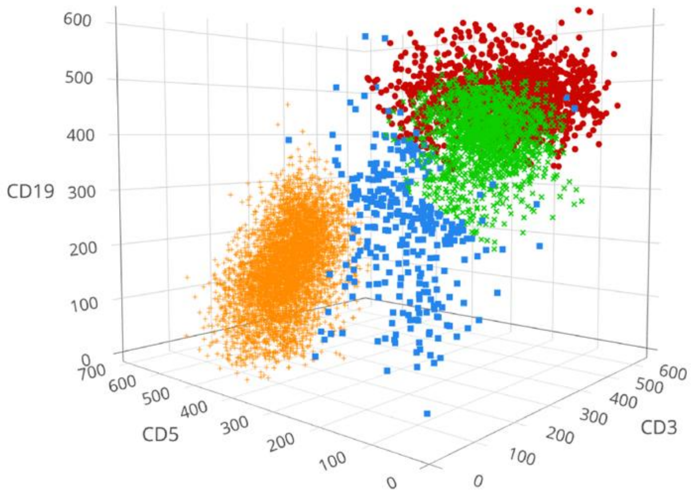

Fig. 1. **Example output for K-means clustering [7]**

PCA (Principal Component Analysis) is a dimensionality reduction technique that transforms data into a new coordinate system, where the greatest variance lies along the first axis, the second greatest variance along the second axis, and so on. It helps reduce the number of features while retaining as much variability as possible.

PCA is a versatile technique that not only reduces dimensionality but also enhances data visualization and interpretation. By identifying the most significant components of variance, it allows for a clearer understanding of the underlying structure of complex datasets, aiding in more efficient decision-making and analysis.

Figure 2 shows an example of PCA, illustrating how the data is transformed into a new coordinate system [8], reducing its dimensionality while retaining the maximum variance.

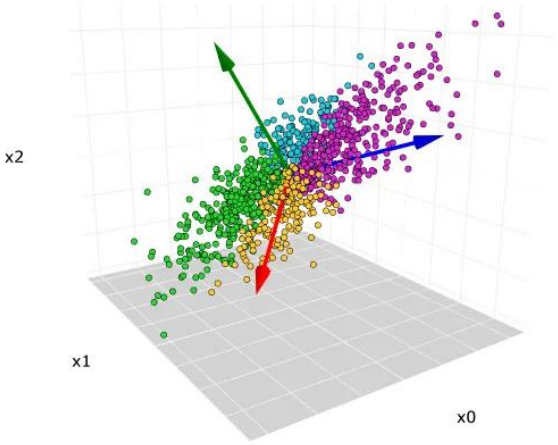

Fig. 2. **Example of Principal Component Analysis [9]**

PCA is used in biology to analyze gene expression data, helping to reduce the complexity of high-dimensional datasets and identify patterns or relationships between genes. PCA was applied to a dataset [10], replicating the separation of brain tissues, cell lines, and hematopoietic tissues from all others in the first three principal components (PCs), as reported by Lukk et al. However, the analysis revealed different orientations of these three PCs. The fourth PC is associated with liver and hepatocellular carcinoma samples, in stark contrast to the noise association reported by Lukk et al.

Figure 3 shows PCA-based analysis of the large-scale structure in gene expression data, with the classification into 7 color-coded groups. The sample distribution among these groups differs, illustrating the variability introduced by different sample compositions.

In terms of liver-specificity, the detection of a liver-specific signal in PC 4 is critically dependent on the number of liver (cancer) samples included in the analyzed dataset. A reduction in the number of liver (cancer) samples to 50 or 60% completely erases any liver-specificity in PC 4. This highlights how the composition of the dataset impacts the PCA results, as the observed difference in PC 4 between the Lukk dataset and used in the study dataset can be attributed to the higher proportion of liver (cancer) samples in our dataset. In the Lukk dataset, the proportion of liver (cancer) samples is only 30% of that in second dataset, explaining the observed contrast in the liver-associated signal.

Thus, the number of liver samples included in the dataset significantly influences the identification of liver-related patterns in the PCA, with smaller proportions of liver samples potentially masking the liver-specific signal in higher PCs.

In conclusion, this thesis explored the concepts of clustering and PCA, highlighting their significance in data analysis. Clustering helps group similar data points, while PCA serves as a powerful dimensionality reduction tool, simplifying complex datasets while preserving key patterns. An example from biology demonstrated how PCA can be applied to gene expression data, revealing meaningful insights into the structure of tissue samples, such as brain tissues, cell lines, and liver cancers. By understanding and applying these techniques, we can uncover underlying relationships in biological data, ultimately advancing research and diagnostics.

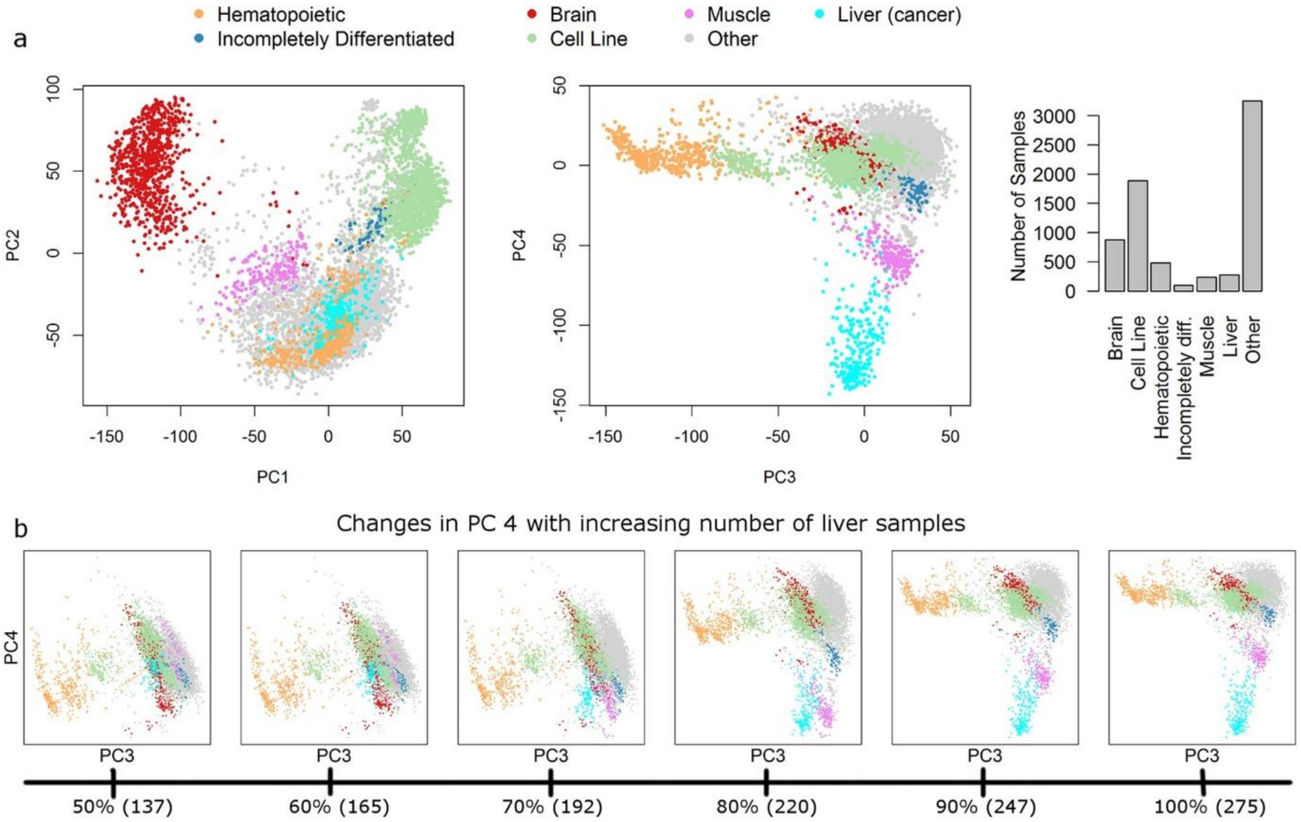

Fig. 3. **PCA based analysis of the large scale structure in gene expression data [10]**

References:

1. What is clustering? URL: https://developers.google.com/machine-learning/clustering/overview#:~:text=Clustering%20is%20an%20unsupervised%20machine,of%20grouping%20is%20called%20classification

2. What is principle component analysis (PCA)? URL: https://www.ibm.com/think/topics/principal-component-analysis#:~:text=Principal%20component%20analysis%2C%20or%20PCA,of%20variables%2C%20called%20principal%20components

3. Kmeans URL: https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

4. Euclidean Distance Formula URL: https://www.cuemath.com/euclidean-distance-formula/

5. Cluster Centroid URL: https://www.sciencedirect.com/topics/computer-science/cluster-centroid

6. What is a data point? URL: https://www.lenovo.com/us/en/glossary/data-points/#:~:text=A%20data%20point%20is%20a,text%2C%20or%20even%20an%20image

7. Identify target user segments with ML Clustering & Classification. URL: https://bookdown.org/travis/data_science_for_human_centered_product_design/Project2.html

8. Coordinate Systems URL: https://phys.libretexts.org/Bookshelves/University_Physics/Book%3A_Introductory_Physics_-_Building_Models_to_Describe_Our_World_(Martin_Neary_Rinaldo_and_Woodman)/25%3A_Vectors/25.01%3A_Coordinate_Systems

9. Principal Component Analysis (PCA) Explained Visually with Zero Math. URL: https://towardsdatascience.com/principal-component-analysis-pca-explained-visually-with-zero-math-1cbf392b9e7d/

10. PCA based analysis of the large scale structure in gene expression data. URL: https://www.nature.com/articles/srep25696/figures/1