National Research University
«Higher School of Economics»
Faculty of Computer Science
Educational programme HSE University and University of London Double Degree Programme in Data Science and Business Analytics
Undergraduate
01.03.02 Applied Mathematics and Computer Science

Report

On Academic Internship Results

in Faculty of Computer Science, Higher School of Economics
(name of organization, enterprise)

Fulfilled by the student of the group #_____

_____
(Surname, First name, Patronymic, if any)

_____
(Signature)

Faculty's Internship Supervisor:

Rudakov Kirill Alexandrovich
Surname, First name, Patronymic, if any

Big Data and Information Retrieval School, Guest Lecturer
CS Faculty Department, Position, Academic title if any

(Date of the check)	*(Grade according to 10-point scale)*	*(Signature)*

Internship Supervisor:

Chernyshev Vsevolod Leonidovich
Surname, First name, Patronymic, if any

Big Data and Information Retrieval School, Associate Professor

(Date of the check)	*(Grade according to 10-point scale)*	*(Signature)*

Moscow, 2024

Table of content

Introduction

The results of exploratory data analysis still remain to be crucially significant for the success of the algorithm developed for machine learning. During this process useful insights are brought up for the programmers which help them to find an individualistic approach during the development process.

One of the most important techniques is dimensionality reduction. It helps to represent the information about some dataset with many features with more convenient and compact components. The most popular and standard approach for dimensionality reduction problem is the Principal Component Analysis. However, the Uniform Manifold Approximation and Project approach method is recently becoming more studied. It proves to preserve both the associations between components in higher dimensions and fast execution speed during its performance.

This report covers the process and results that were attained during the work experience internship dedicated to the study of the UMAP method which was offered by the associate professor Chernyshev Vsevolod.

Object and purpose

The identified objectives for the successful completion of the course are:

1. Study the article.
2. Implement the code.
3. Conduct an analysis.
4. Prepare the report.

The purpose and desired result after the internship is the understanding of the applicability of the UMAP dimensionality reduction algorithm.

Calendar Schedule

№	**Calendar period**	**Plan of work**[1]	**Supervisor's mark on the point fulfilment (signature)**
1	15.07.2024	1. Instruction on familiarization with the requirements of labor protection, safety, fire safety and internal labor regulations	
2	15 - 31.07.2024	2. Study of the article and correction of the mentioned case	

[1]*For all students, the **first point of Plan of Work** (Instructing on the requirements) is a standing point and **could not be modified**. Please, do not edit it. **All other points are sample points** - you can change it according to your Individual Internship Assignment.*

3	31.07.2024 - 20.08.2024	3. Studies in the estimation of the proportion of the explained variance; preparation and submission of the report	

Description of explored medium

The exploration of the UMAP method was divided into two parts: study the overall performance principles and conduct an analysis of estimation of explained variance based on comparison with the PCA method. In order to accomplish the mentioned steps several resources were used during the internship.

The description and visual representation presented in [1] provides a slow yet meaningful introduction to the overall conception of the studied method. The basic idea of UMAP can be summarised in the fact that it utilises the construction of high-dimensional graphs in order to optimise and correct the low-dimensional one with the help of similarity scores between clusters of neighbours. The algorithm is designed in such a way that helps to preserve both the local and global structure of items and dataset. The influence of different hyperparameters was tested and analysed by using the possibilities provided in the referenced material.

For the investigation in the field of approximating the explained variance obtained from UMAP components the material in [2] written by Oskolkov Nikolay was proposed for study by the internship supervisor. The article firstly displays the relevance of studying UMAP. By applying the algorithm to the MNIST dataset it is discovered that results obtained turn out to be much more convenient for use than the ones obtained from the standard PCA. Subsequently, by performing derivations from findings discovered with the principal components method, the work dives deeper into the investigation of possible ways to estimate the ability of each component of the studied method to explain the total variance in the given data.

Description of achieved results

During the process of getting acquainted with the principles and underlying mechanisms of UMAP it was discovered that the two most important hyperparameters of the algorithm are "*n_neighbors*" and "*min_dist*". Their meaning and usage helps to understand the overall performance of the method. The first one controls the number of neighbours each point in a high-dimensional graph must have in order to be included into some cluster. The second hyperparameter sets the minimal distance between the points in the low-dimensional space. Both of them help to maintain the balance between the

formation of local and global structures. Moreover, the values these hyperparameters possess influence the calculation of the similarity between the clusters and further construction of the low-dimensional picture.

For the purpose of getting practical examples of how well UMAP performs on the real data an individual project in Jupyter Notebook was created. The results of experiments in this and further paragraphs can be checked by accessing the resource [3]. The performance of the algorithm was compared with PCA. Besides the fact that this method allows to generally outshine the advantages of the studied method, it is also needed in order to imply the assumptions needed for latter explorations.

The two algorithms were directly applied to the MNIST dataset. The comparison of performance is done by visualising the results on 2- and 3-dimensional planes. Firstly, the PCA method is applied and plotted on the graphs. The results of the algorithm can be viewed on Figure (1).

Figure 1. Results obtained from the application of PCA to the MNIST dataset.

As it can be observed the plots display complete lack of ability to provide sufficient information about the dataset. So, the algorithm was then changed to UMAP, applied to the same dataset and visualised in the same dimensions. Plots received from these actions are represented on Figure (2). Performing visual analysis allows to make a conclusion that the results obtained from the application of UMAP give the possibility the retract much more informative details about the dataset.

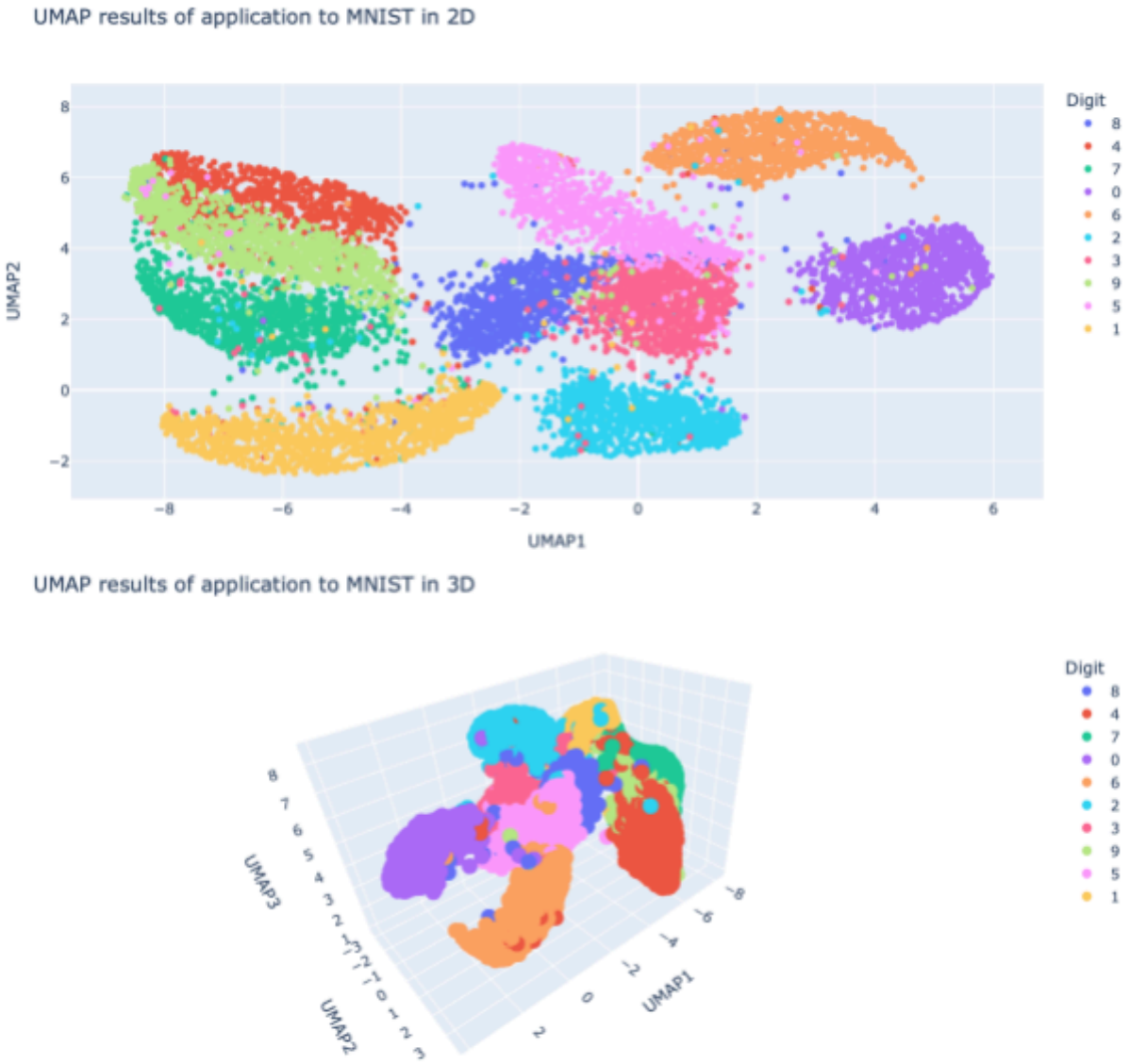

Figure 2. Results obtained from the application of UMAP to the MNIST dataset.

One important note must be mentioned. During the hyperparameter tuning, "*init*" is needed to be set to initialise the initial positions of features in low-dimensional space. Its value must be based on the results obtained from application of PCA with the optimal number of principal components. This can be discovered by plotting the amount of cumulative variance explained with each additional component (displayed on the Figure (3)). Around 90% of the total variance of the

dataset can be explained by reducing the number of features to 60-65. So, the statement that the optimal number of components being equal to 62 made in [2] was verified.

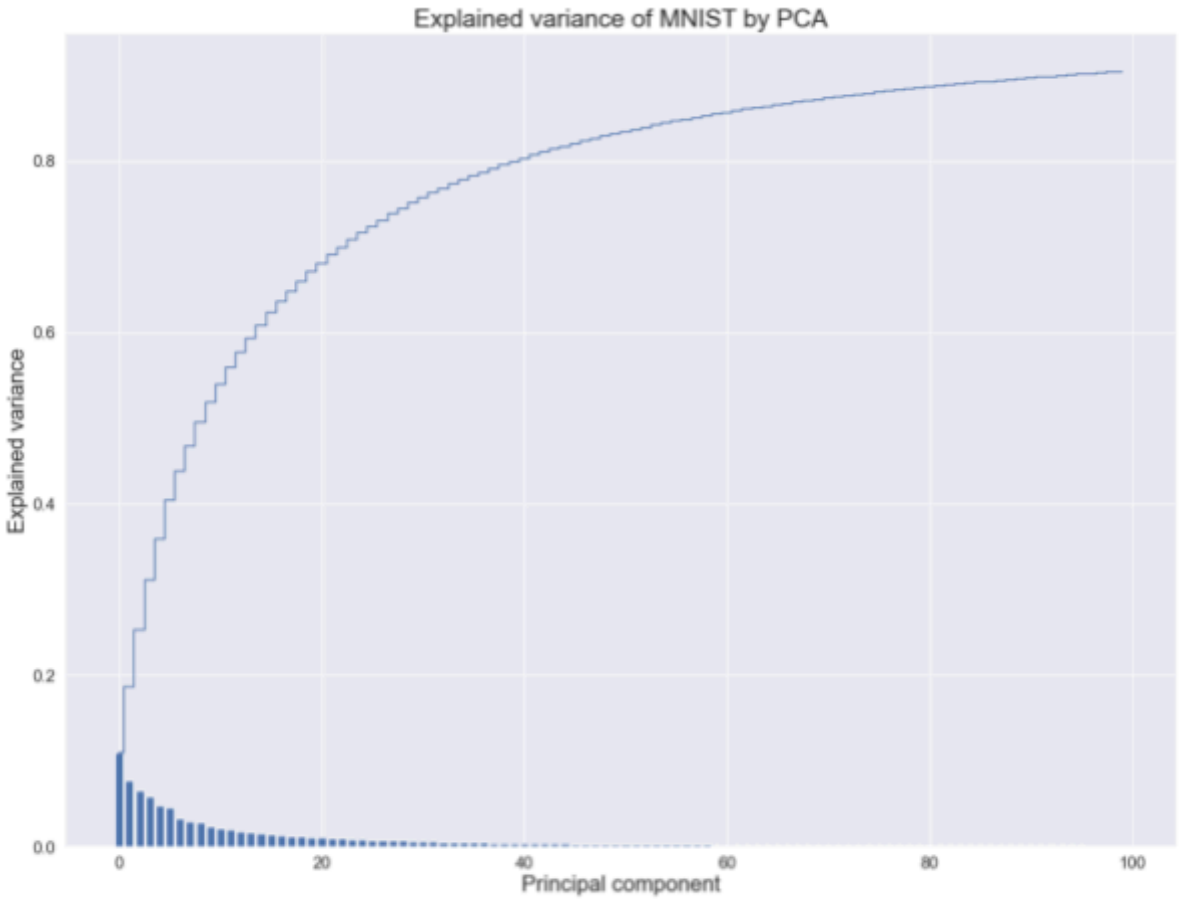

Figure 3. Proportion of explained variance by the components of PCA.

Finally, the work deepens in the investigation of possibility to estimate the proportion of explained variance by each component of UMAP. Such topic possesses high interest due to the fact that the non-linearity of the studied method prevents from applying the standard procedures.

The repeated usage of the PCA method is justified because of the fact that the knowledge of how to calculate the proportion of explained variance from each component allows to make conclusions about the suitability of applied approximation. So, by calculating the coefficient of determination obtained from the application of the partial least squares procedure the possibility of making such a substitution was verified. The results tested on PCA showed close to identical approximation of the proportion of explained variance. The obtained results imply that such a procedure is suitable to be applied for the UMAP method. Following this assumption, the approximated values of explained variance for the first three components of two mentioned methods were calculated and can be observed plotted on the Figure (4).

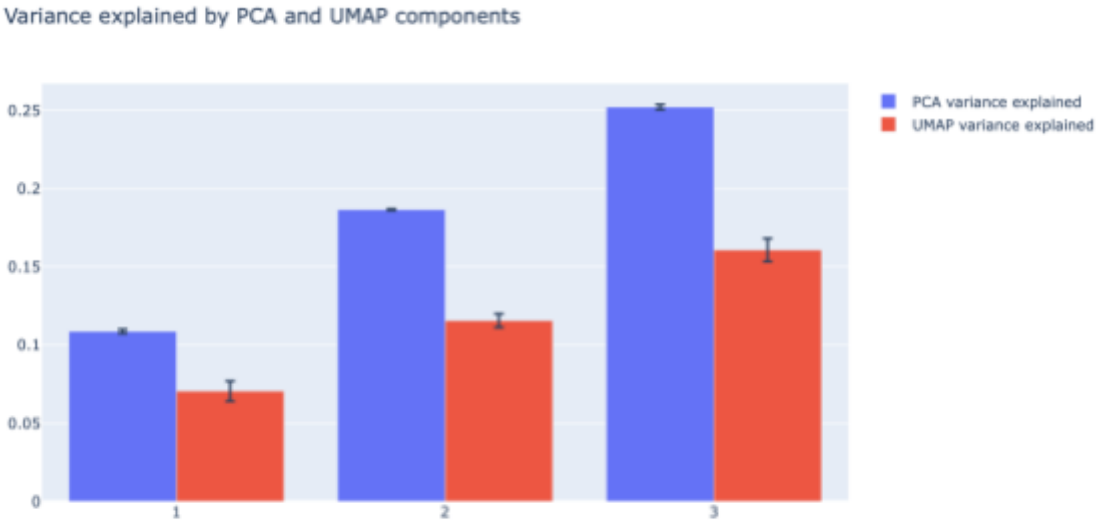

Figure 4. Proportion of explained variance by the components of PCA and UMAP approximated by PLS.

The visual analysis of the plot allows to make another supporting argument for the approximation as the goal of PCA is to select those components that maximise the amount of variance explained. However, UMAP is not aimed at this goal and has noticeably less values compared to standard principle components. Furthermore, it appears that the confidence interval for the approximation of explained variance in components of the studied method is also wider. Such a detail signals that despite the fact of having reasonable standard error now results derived from PLS need to be studied with additional attention.

Nonetheless, an intriguing detail was discovered. Having less proportion of explained variance per component, UMAP still manages to outperform PCA given the results described in the beginning of this section. The final step of the research in the internship was to verify the assumption that the actual solution to the problem is that the linkage between the components of UMAP and the labels of MNIST dataset is much stronger than the ones with the linear method.

Such an implication was firstly checked with the calculation of coefficient of determination from PLS regression and by construction of a correlation heatmap. Both of these actions verified the assumption. However, as it was mentioned, the results related to the approximation of UMAP explained variance by using the partial least squares method needs to be studied more thoroughly. As a result another plot depicting the amount of explained variance in labels with the confidence intervals was done. The analysation of results on Figure (5) suggests that the implication of linkage between the UMAP components and MNIST labels might not hold true. This conclusion is made due to the fact that confidence intervals of some components are wide and include the values that standard principal components might possess. Thus, the robustness of the assumption must be studied in more detail.

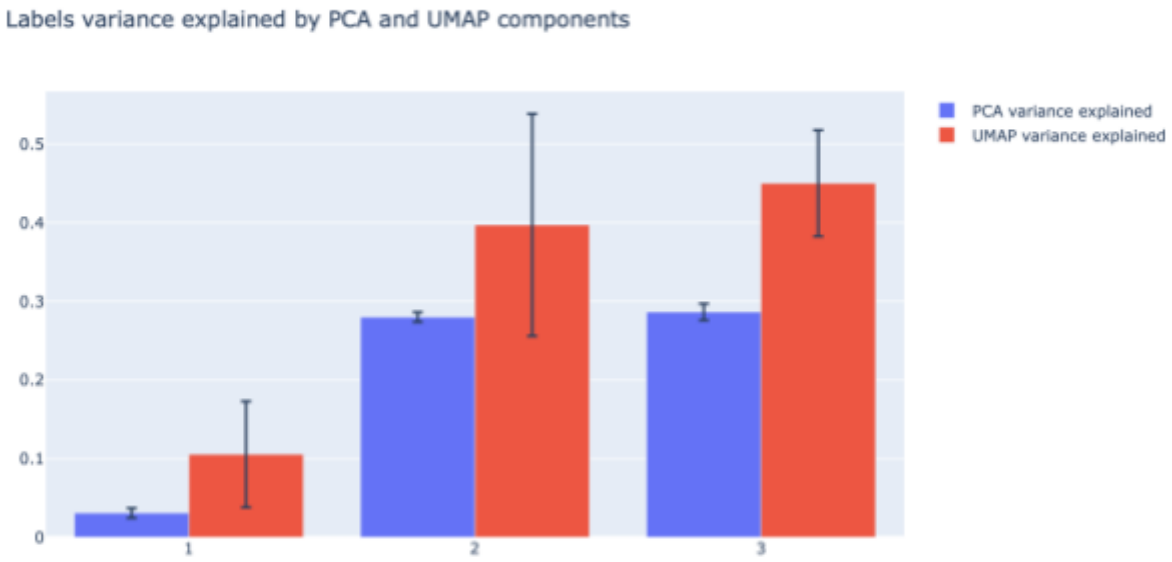

Figure 5. Proportion of explained variance in labels by the components of PCA and UMAP.

Conclusion (main results and findings)

After the completion of the internship the following results were obtained:

- The study of the UMAP method was performed. Its main ideas and performance were regarded and analysed.
- Comparison of UMAP with PCA was performed. The advantages of UMAP were observed on the MNIST dataset example.
- The possibility of estimating the amount of explained variance by each component of UMAP was studied.

Overall, the internship provided great possibility to study the state of the art technique in the world of machine learning. The UMAP method provides an alternative possibility to view the advantages of dimensionality reduction. The advantages of the method compared to the other algorithms outshine UMAP and prove it to be prospective.

In conclusion, the analysed material during the internship broadens the knowledge in the sphere of machine learning techniques. The studied method gives the ability to analyse datasets from another point of view. Yet the exploration of the phenomenon discovered during the work leaves great prospects for further investigations in how the algorithm is structured.

The List of Sources

1. **Understanding UMAP // Andy Coenen, Adam Pearce. Access mode: https://pair-code.github.io/understanding-umap/ (date of application: 19.08.2024).**

2. **UMAP Variance Explained // Nikolay Oskolkov. Access mode: https://towardsdatascience.com/umap-variance-explained-b0eacb5b0801 (date of application: 20.08.2024)**

3. **UMAP_Study // Gorelskii Roman. Access mode: https://github.com/RomanGorelsky/UMAP_Study.git (date of application: 20.08.2024)**

Appendix

Grading Scheme

ECTS Grades	10-point scale (Exam)	5-point scale
A+	10,00	Excellent
A	9,00	Very good
A –	8,00	Very good
B+	7,00	Good
B –	6,00	Good
C +	5,00	Satisfactory
C –	4,00	Satisfactory
F	3,00	Fail
F	2,00	Fail
F	1,00	Fail