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 popular variations of such methods are the standard Principal Component Analysis and a slower non-linearly advanced version called t-Stochastic Neighbour Embedding. However, the Uniform Manifold Approximation and Project approach method is recently becoming more studied. It proves to both preserve the ability to maintain associations between components in higher dimensions and work much faster than tSNE.

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 reporte	

Description of explored medium

The exploration of the UMAP method was divided into two parts: study the overall performance principles and compare them with other already existing algorithms, namely PCA and tSNE. 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 comparison of three methods of dimensionality reduction the material in [2] written by Oskolkov Nikolay was proposed for study by the internship supervisor. The article discusses the advantages of UMAP compared to PCA and tSNE in performance. By running the tests on MNIST dataset the conclusions were derived that the non-linear approach allows distinguishing the handwritten digits much better than the linear one used in principal components. Moreover, UMAP's ability to perform calculations much faster and obtain the same results throughout a series of experiments prioritises its use over the tSNE algorithm.

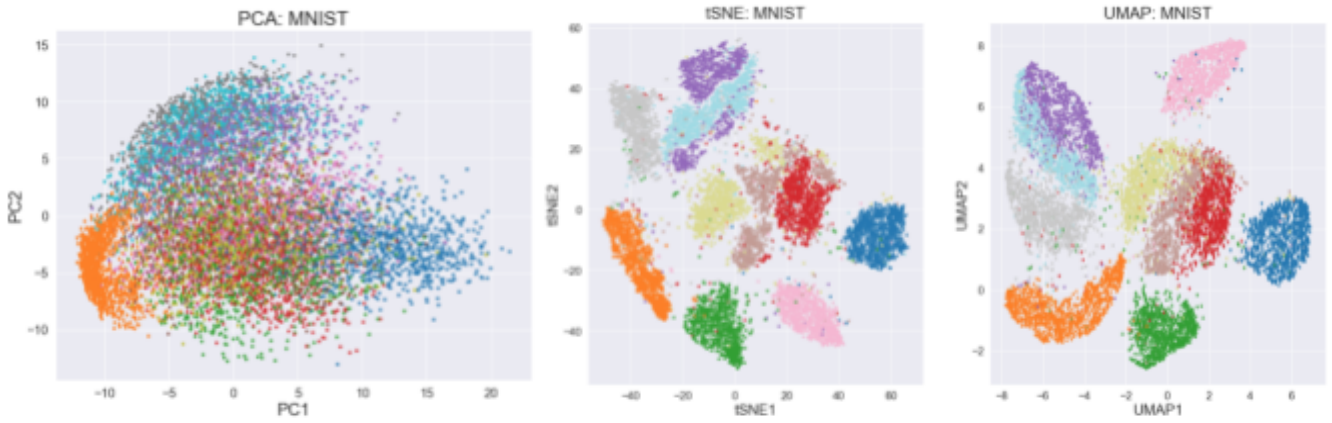
Description of achieved results

During the process of examination of resource [1] 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 UMAP. 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.

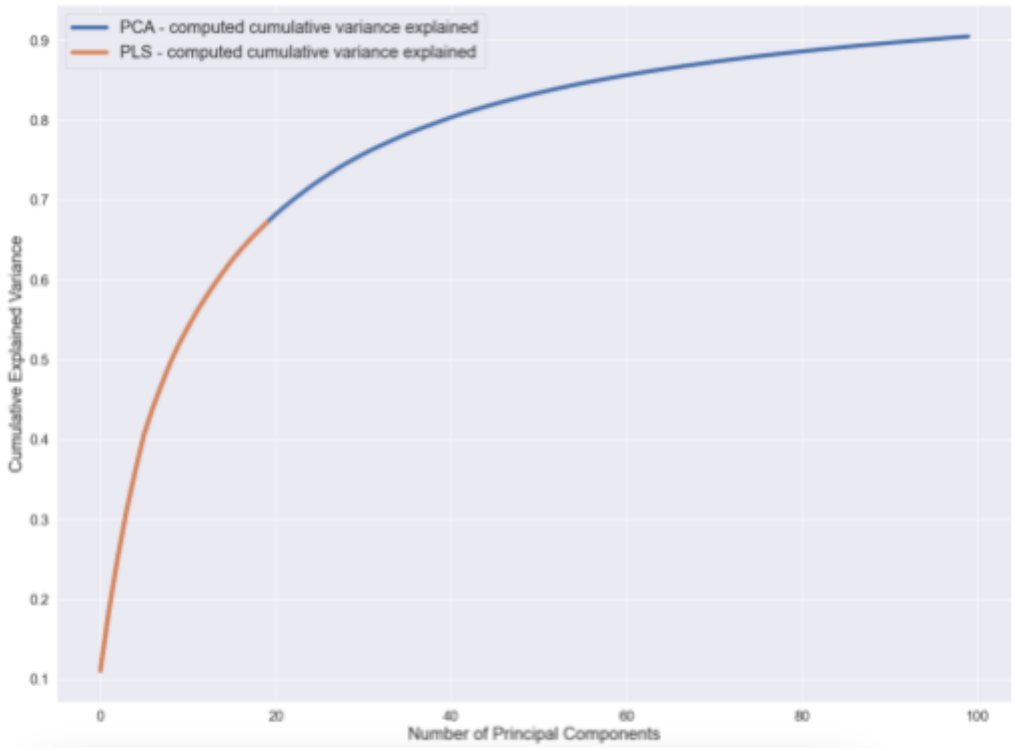
The second part of the internship was able to be done with the use of Jupyter Notebook. The code presented in [2] was transmitted into the notebook. During the work it was discovered that some parts of the code needed to be updated for newer versions of Python and used libraries. All needed modifications were performed and the final version of the notebook is displayed at [3].

The overall purpose of the second part is designed to get a grasp of how UMAP outperforms PCA and tSNE method on the real data. It is done by directly applying the three methods on the MNIST dataset which is prepared beforehand for the task. The visualisation of the dataset using the two components from every algorithm allows us to acknowledge the superiority of UMAP compared to the other two. This is achieved due to several features which UMAP possesses. The first one is that the dimensionality reduction technique used in it is non-linear compared to the linear matrix factorisation applied in PCA. Secondly, the initial low-dimensional graph created by the algorithm is initialised non-randomly as opposed to tSNE. This concludes that the results of UMAP are always stable and can be replicated numerously.



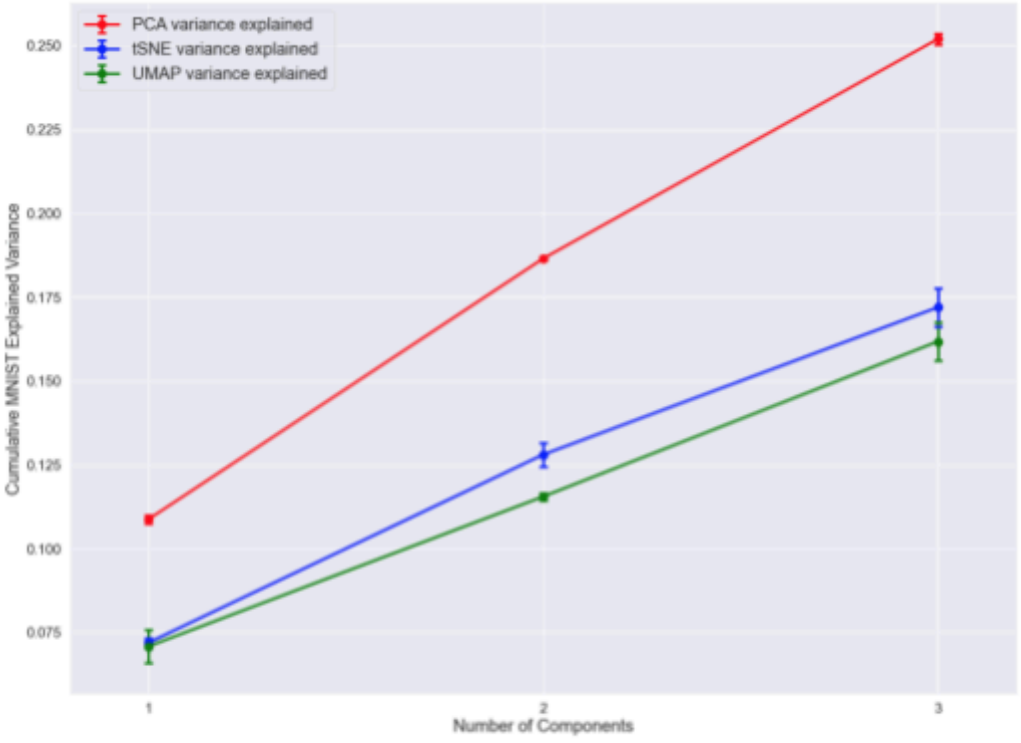
Picture 1. Visualisation of the MNIST dataset (coloured for the distinction of digits) using the first two components of every algorithm.

Furthermore, the interest falls on the exploration of finding a way to compute the percentage of explained variance by components in UMAP as it is done with PCA. However, due to the non-linearity foundation of the algorithm it cannot be calculated directly. The approximation is done by utilising the coefficient of determination obtained from the partial least squares procedure. It is firstly verified on the PCA results. The performed test depicts that the approximation and the true values of explained variance coincide quite well. It is depicted in picture (2).



Picture 2. Cumulative variance explained from PCA compared to the results obtained from PLS regression.

The derived results transmit the following assumptions on the rest two methods. The PLS procedure was then applied to tSNE and UMAP. The cumulative variation explained by the first three components was then plotted with the results obtained previously from PCA.

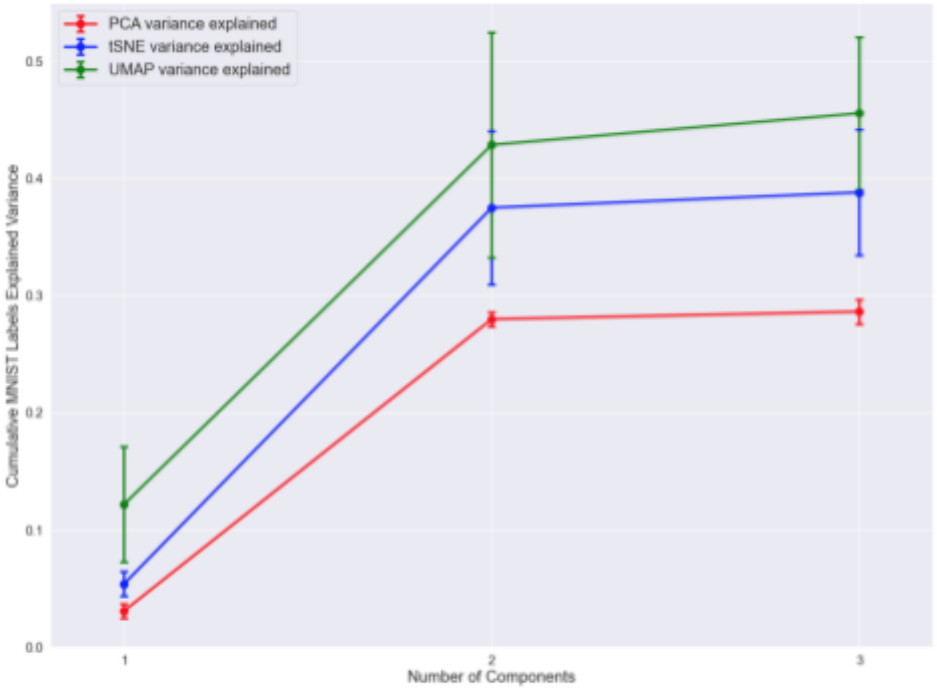


Picture 3. Cumulative variance explained by first three components from PCA (red), tSNE (blue) and UMAP (green) obtained from PLS procedure.

The depicted values reassure that the application of PLS for estimation of the variance explained by UMAP components is sufficient. From the visualisation it can be concluded that UMAP and tSNE do not seek to maximise the variance explained by each component in contrast to PCA.

However, the final results of the non-linear methods are drastically better than those obtained from the linear one. This pushes the research further to investigate the so-called “biological” variation explained by each algorithm. It is studied with the introduction of a hypothesis that the components because of non-linearity are linked to the labels of the dataset rather than the total variation contained in it. This assumption was verified with the construction of a heatmap and a matrix obtained from the PLS regression.

Final step of the study was to investigate what is the influence of increasing the number of components on the results of dimensionality reduction techniques. The tests were performed with all three algorithms. The results were plotted once again depicting the cumulative labels explained variance.



Picture 4. Cumulative labels variance explained by first three components from PCA (red), tSNE (blue) and UMAP (green) obtained from PLS procedure.

As a result, leading components of UMAP capture much more of label variance than the components of PCA. This conclusion is counter intuitive following the conclusions made from picture (3). So, this effect of linkage to the labels with less explanation of total variation in pixels of the dataset leaves room for further investigations.

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 and tSNE 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.

It is important to mention that during the replication of the experiments presented in [2] some of the results could not be obtained as they were shown in the article. Firstly, the results from the calculation of coefficient of determination proposed through the formula differ from those presented by the author. Nevertheless, the verification that PLS method is a good approximation of explained variance is still done due to the construction presented in picture (2). Secondly, the confidence intervals obtained for the results plotted on picture (4) are a lot wider than the ones shown in the article. However, the overall conclusion made from this plot by the author still holds.

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. https://pair-code.github.io/understanding-umap/
2. https://towardsdatascience.com/umap-variance-explained-b0eacb5b0801
3. https://github.com/RomanGorelsky/UMAP_Study.git

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