Machine Learning - Homework 8 Report

Lukas Johannes Ruettgers (2023403372)

December 1, 2023

1 Introduction

Visual or lingual data is often measured as high-dimensional feature vectors. Since the expressivity of the feature space grows exponentially with the number of features, many well-studied learning algorithms do not scale well with the dimensionality of the data. For that reason, feature extraction seeks to find a low-dimensional representation that still accurately approximates the original data distribution. To that end, many dimensionality reduction algorithms have been devised, which mostly fall into two categories.

1.1 Principal Component Analysis

The first category comprises matrix factorization algorithms, which try to decompose the original data matrix $\mathbf{X}$ into simpler matrices. A popular method is the *Principal Component Analysis* (PCA), whose objective is to determine an eigenvector basis of the covariance matrix of $\mathbf{X}$. This eigenvector basis constitutes an orthogonal matrix, which rotates the feature spaces to their principal axes. The hidden features along these principal axes are uncorrelated, which effectively diagonalizes the covariance matrix. To obtain an efficient trade-off between dimensionality and approximation of the original data, the eigenvectors can be ordered by the variance of their corresponding hidden features. Finally, the number of the most influential hidden features to which the high-dimensional data shall be reduced is determined by the desired fraction of the variance one wishes to preserve.

1.2 Uniform Manifold Approximation and Projection

Matrix factorization algorithms can naturally only express linear transformations of the original data and are hence limited when the topology of the high-dimensional data does not follow linear patterns. For that reason, graph layout algorithms seek to describe the original data in form of a graph, whose connectivity should be related to the proximity of vectors in the high-dimensional manifold. The problem then shifts to obtain a low-dimensional graph that optimally captures the complex edge relations of the high-dimensional graph. In *Uniform Manifold Approximation and Projection* (UMAP), this high-dimensional graph is broadly speaking constructed by locally approximating the topology around each data point by an isotropic sphere, whose size extends just to the $\mathrm{N}_{\mathrm{Neighbors}}$th nearest neighbour of that point. Any points whose spheres overlap will be connected in the high-dimensional graph by an edge, whose weight decays exponentially with the distance of their corresponding points. This requires both a number of neighbours $\mathrm{N}_{\mathrm{Neighbors}}$ such as a distance metric to be specified in advance. The resulting weight distribution can be regarded as defining a fuzzy set around each point. Finally, we avail to well-known optimization algorithms to obtain a low-dimensional topology that minimizes the cross entropy between the distance-based membership functions of the fuzzy sets in the high-dimensional and low-dimensional topology.

2 Experiment Objective

The following experiment compares the low-dimensional embeddings obtained by both PCA and UMAP. As high-dimensional dataset, we use the *MNIST Digits* dataset, which contains 60,000 grayscale 28×28 images of handdrawn digits. While we investigate possible meanings of the principal components obtained by PCA on the digit 5, we will study the influence of the hyperparameters on the UMAP embedding to better understand their importance and sensitivity. Although the MNIST Digits dataset exhibits the accommodating property that each feature shares the same domain of greyscale values in $\{0, \ldots, 255\}$, this does surely not imply that the distribution of greyscale values across different features are the same. In

particular, the pixels at the border of each image will surely be black, while pixels in the middle right are likely to have high (white) values because most of the digits typical shape contains strokes in this region. Applying the Euclidean distance metric per default seems therefore a courteous choice, since the Euclidean metric will weigh distance along each principal axes equally. For that reason, one might desire to cancel out the individual variances by using the Mahalanobis distance. In our case where the covariance matrix is diagonalized, the Mahalanobis distance would effectively normalize each dimension by its variance. This motivates investigating whether the variance-normalized Mahalanobis distance achieves better separation than the default Euclidean metric.

Specifically, we will in total consider the following hyperparameters:

- $N_{\text{Neighbors}}$: Number of neighbours to which each sphere in the high-dimensional manifold should be extended.
- MINDIST: Minimum expected distance between the points in the low-dimensional embedding.
- SPREAD: Spreads the embedded points away from the origin.
- Distance metric: We will compare the default Euclidean metric with other family members of the Minkowski metrics ($p = 3, 5$) and the Mahalanobis distance.

Even though MINDIST and SPREAD do not improve the cluster separability in the embedding but only affect the visualization, it is important to understand how they contribute to an adequate visualization. In particular, the documentation does not provide information on the sensitivity of the visualization quality to the hyperparameter SPREAD.

3 Results

3.1 Principal Component Expressivity

To express 80% of the variance in the original data, one requires only the first 45 principal components (PCs) of the eigenvector basis obtained by PCA. The ratio of unexplained variance decreases exponentially and the evolution of its counterpart – the ratio of explained variance – is depicted in Fig. 1. As the plot describes, the first two PCs account for approximately one sixth of the variance, while the first and the third PC together express one to two percent less.

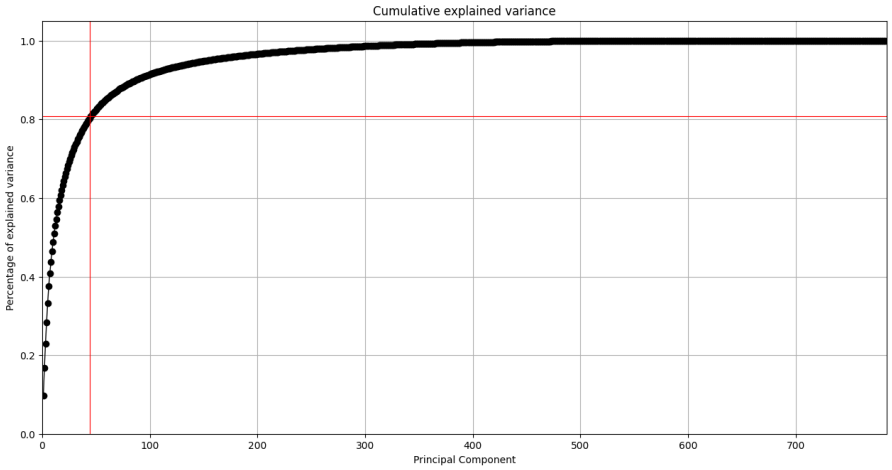
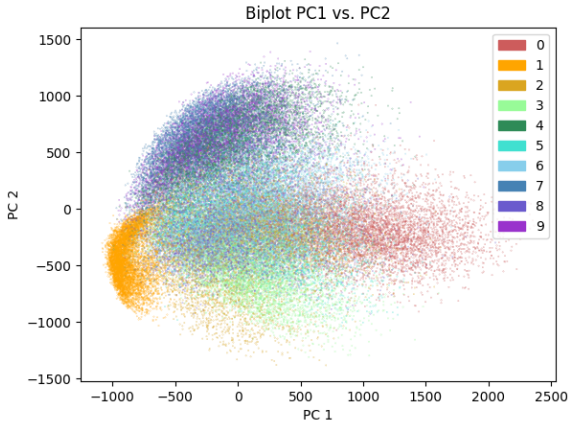
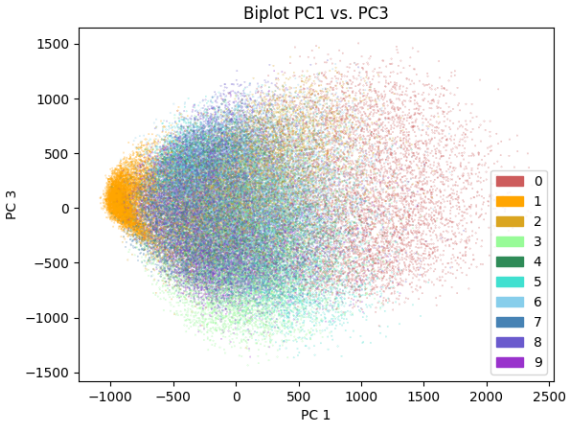


Figure 1: Proportion of cumulative variance explained by the Principal Components.

When we restricts ourselves to only two PCs, the comparison between the feature distribution of the first and the second PC in Fig. 2a and the first and the third PC (Fig. 2b) indicates that the second PC exhibits some ability to tell the digits $7, 8, 9$ apart from the remaining digits. The feature distribution of the third PC alone seems not to provide a useful discrimination, which is not surprising because it can only represent approximately six percent of the entire variance. On the contrary, the first PC manages to strongly discriminate 0s from 1s.



(a) First vs. second PC

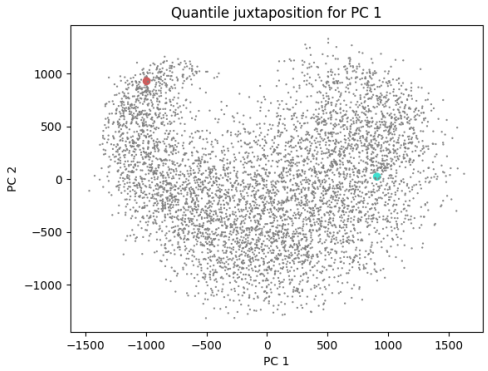


(b) First vs. third PC

Figure 2: Projection of MNIST Digits to the subspaces spanned by the most important Principal Components respectively. While still exhibiting overlap, the PC values for each digit seem to center around a specific region.

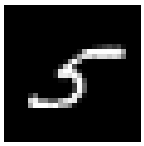
3.2 Quantile Comparison

If we limit ourselves to the images that represent the digit 5, we obtain new first and second PCs. The distribution of the digit images along these two hidden features is shown by the grey point clouds in Fig. 3.

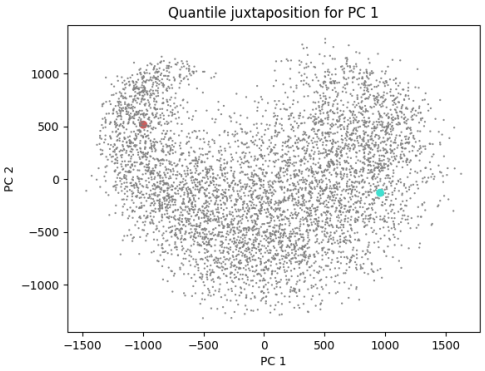


(a) PC1 quantile sample 1

(b) $q_{0.9}$

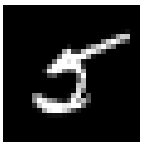


(c) $q_{0.1}$

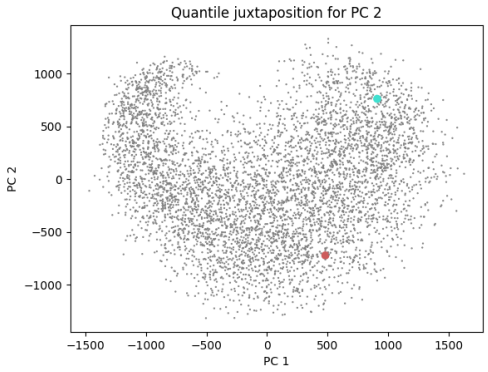


(d) PC1 quantile sample 2

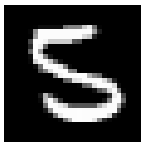
(e) $q_{0.9}$



(f) $q_{0.1}$

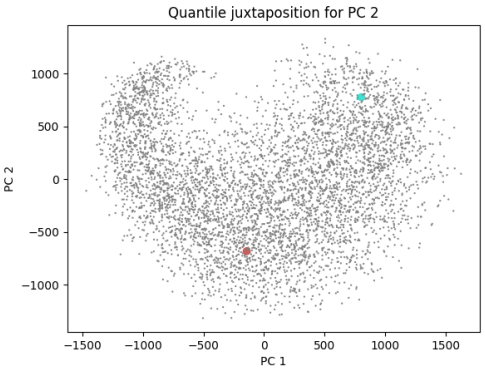


(g) PC2 quantile sample 1



(h) $q_{0.9}$

(i) $q_{0.1}$



(j) PC2 quantile sample 2

(k) $q_{0.9}$

(l) $q_{0.1}$

Figure 3: Juxtaposing high and low quantiles along the most important Principal Components to unravel the meaning behind the obtained hidden features. In the biplot, the 0.1 quantile $q_{0.1}$ is marked by the red dot, while the turquoise dot indicates the 0.9 quantile $q_{0.9}$. The gray point cloud represents the rest of the digit 5 images. Only the highlighted points differ across images, the underlying point cloud is the same.

To get an impression of the meaning the first two PCs could bear, 1000 of these 5 digit images were sampled and for each PC, the 0.1 and 0.9 quantiles $q_{0.1}, q_{0.9}$ were computed to estimate an representative for images with quite high and small values. For each such quantile, five samples have been computed. For example, two 0.9 quantiles of the first PC are depicted in Figures 3b and 3e and their position in the distribution of the first two PCs is highlighted in Figures 3a and 3d.

A comparison with the low quantiles in Figures 3c and 3f quickly reveals that digits in the low PC1 region have been written with a strong horizontal distortion towards the right. In particular, the top line of the 5 goes far to the right and almost hits the border in the most extreme cases.

Conversely, the two images representing the high quantiles are more difficult to categorize. On the one hand, the digit in Fig. 3b is carefully drawn, where the intersection of the upper two lines form a right angle and the half circle in the bottom part remains steady in its radius. This description matches with a large proportion of the sampled high quantiles of PC1. On the other hand, the digit in Fig. 3e does not meet this description at all. It is carelessly drawn and rather resembles the letter 'S'. Moreover, its line thickness is relatively large. The main attribute that these digits share in common is that they do not meet the characteristics of the low quantiles, meaning that they are not distorted horizontically and instead placed at the center of the image.

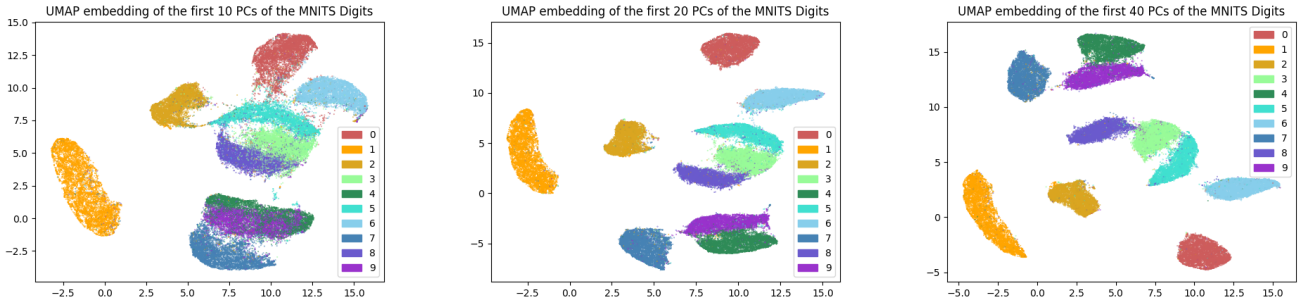
What is notable about the low quantiles of the second PC in Figures 3i and 3l is that the angle of intersection between the upper vertical line and the bottom three-quarter circle is quite steep and the corner lengthened more than usual. These characteristics do not fully apply for the high quantiles of the second PC. For example, the vertical line of the high quantile in Fig. 3h is already turned so much to the right that it directly transitions into the circle. While the other high quantile in Fig. 3k exhibits a right angle between the vertical line and the curve, its intersection is not as extremely lengthened as for the low quantiles.

Of course, the true meaning behind the principal components is hard to grasp since they firstly correspond to the concept of a rotation in the original feature space, which might be hardly tractable for the human way of visual thinking. Secondly, approximation errors are innate to each PC, since it represents only a small fraction of the data's entire variance.

3.3 UMAP

3.3.1 PCA embedding for UMAP

This limitation of the first PCs is also represented by the series of embeddings in Fig. 4. When providing UMAP solely the first ten PCs, the clusters in the low-dimensional representation are still relatively noisy. This noise is attributable to the unexplained variance and decreases if we provide UMAP more PCs. Besides noise, the overlap between the cluster for 4 and 9 is still large in the embedding with 10 PCs (Fig. 4a), while these and other clusters become more separable in Fig. 4c.

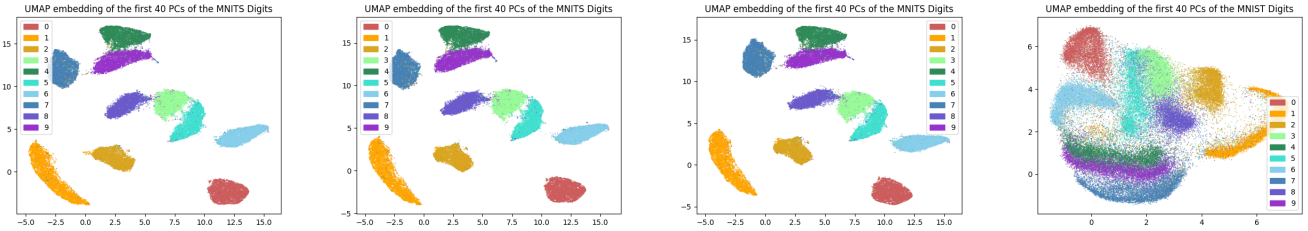


(a) First 10 PCs (b) First 20 PCs (c) First 40 PCs

Figure 4: Applying the UMAP embedding for an increasing number of Principal Components of the dataset, where $N_{\text{Neighbors}} = 15, \text{MINDIST} = 0.1, \text{SPREAD} = 1.0$. While the low-dimensional representation already soundly separates the clusters, the embedding becomes more crisp for an increasing number of Principal Components considered.

3.3.2 Minkowski distance metric

As we observed in Fig. 1, the variances of the first PCs decrease exponentially and distance across different principal axes and are indeed not normalized. One might fear that the Euclidean distance metric for a UMAP embedding of the PCA components will therefore not be a suitable choice, because the membership functions of the fuzzy sets will favour distances along principal axes with low variance over distances along principal axes with high variance.



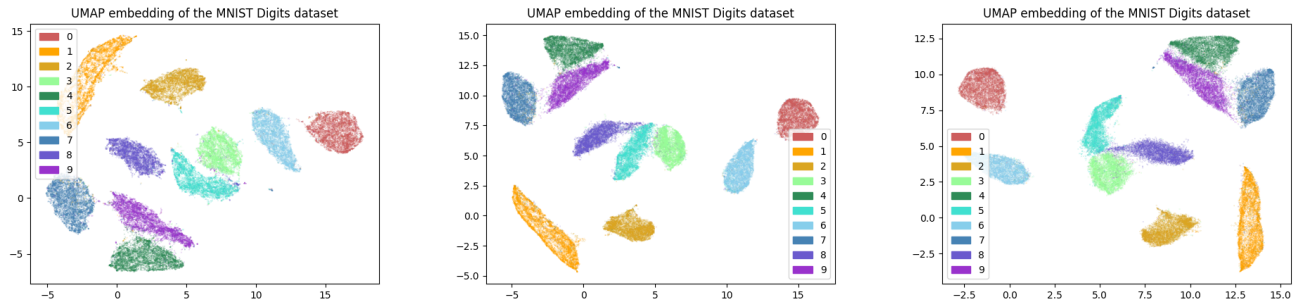
(a) $p = 2$ (b) $p = 3$ (c) $p = 5$ (d) Mahalanobis distance

Figure 5: Comparison of the UMAP embedding using both the Mahalanobis and the Minkowski distance metric for different choices of p. These results were generated with $\mathrm{N}_{\mathrm{Neighbors}} = 15, \mathrm{MINDIST} = 0.1$ and $\mathrm{SPREAD} = 1.0$.

However, when I formulated this hypothesis in the experiment objective, I neglected that this is not a flaw, but conversely an advantage for the embedding of the principal component space, as the juxtaposition in Fig. 5 supports. Recall that in Fig. 2a, the cluster for 0 exhibited by far the highest values for the first PC, while the cluster for 1 clumped at the lowest value region. The large variance across this axis is important to connect to the correct neighbours in the manifold, since the variance of this PC provides an informational gain to discriminate between different clusters. For the normal Euclidean metric (Fig. 5a), the clusters are therefore well separated. On the contrary, this informational advantage is destroyed if we apply the Mahalanobis distance metric, because we normalize by the variances. Different values across inferior PCs then become equally important as different values along superior PCs, which is however not true. Consequently, the clusters are barely separated in Fig. 5d and have been contracted to the origin. This clearly disproves my hypothesis in the experiment objective.

Increasing p inside the family of Minkowski distance metrics increases the tendency by which we favour points with differences along *multiple* PCs over points with differences along few PCs. However, the results for $p = 5$ (Fig. 5c) barely improve the separation already achieved by the Euclidean metric ($p = 2$). One might even say that it slightly aggravates the separation between the clusters 3 and 7, and 4 and 9, respectively. However, this difference is too subtle to infer a clear dominance of the Euclidean metric. However, we can again avail to the argument that the given scale of the variances might get distorted if one does not equally weigh all distances similarly, as the Euclidean does.

3.3.3 Neighbourhood



(a) $\mathrm{N}_{\mathrm{Neighbors}} = 5$ (b) $\mathrm{N}_{\mathrm{Neighbors}} = 15$ (c) $\mathrm{N}_{\mathrm{Neighbors}} = 50$

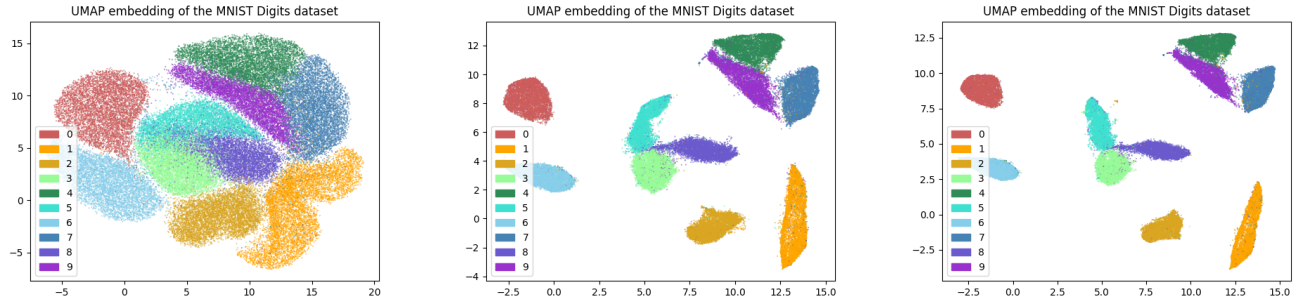
Figure 6: Varying $\mathrm{N}_{\mathrm{Neighbors}}$ for fixed $\mathrm{SPREAD} = 1.0$, $\mathrm{MINDIST} = 0.1$ and the Euclidean distance metric.

Increasing the number of neighbours to which the fuzzy sets in the high-dimensional spaces shall be expanded provides a more accurate global representation of the data at the cost of neglecting the local topology of the manifold. Fig. 6 illustrates this trade-off because the clusters for $3, 5, 8$ and $4, 7, 9$ are gradually moving closer and eventually become connected at their ends. While Fig. 6c provides a better insight which digits might broadly have closer values, Fig. 6a might preserve the most of the local topology, because the hypersphere around each point in which the fuzzy set approximates the manifold with an isotropic membership function becomes smaller.

3.3.4 Minimum distance

Surprisingly, large differences in the minimum distance between the embedded points do not influence the scale at which these points are depicted. It seems that the center of each cluster is not the anchor point from which the cluster is blown up. Instead, this anchor point seems to lie more on the outward of each cluster, causing the cluster to expand more to the center of the plot.

Fig. 7a demonstrates what happens if MINDIST is chosen too large. Then the cluster borders will become quite close, or even fully disappear for the clusters $3, 5$ and 8. As long as one chooses a small minimum distance like in Fig. 7b, the separability between the clusters is ensured. However, this minimum distance must be regarded in relation to the scale of the entire manifold dataset. It might still be possible that other datasets with smaller scales lose their separability already for smaller values of MINDIST.



(a) MINDIST $= 1.0$ (b) MINDIST $= 0.1$ (c) MINDIST $= 0.01$

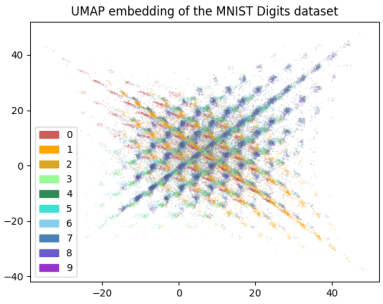
Figure 7: Decreasing the minimum distance between the embedded points for fixed $N_{\text{Neighbors}} = 15$ and SPREAD $= 1.0$ illustrates how the clusters become more contractive. For MINDIST $= 0.001$, the embedding remains similar to Fig. 7c and is hence not depicted here.

3.3.5 Spread adequacy

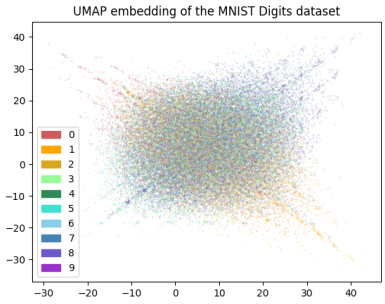
Even after MINDIST is determined adequately, an inadequate choice of SPREAD can still drastically deteriorate the visualization quality (Fig. 8b). Because SPREAD may not be smaller than MINDIST, we chose MINDIST $= 0$ for the experiments, but the results were similar for MINDIST $= 0.01, 0.1$.

While MINDIST seems to expand the clusters towards the center, SPREAD values larger than 1.0 will bloat the clusters up to the outward. The distance from the center grows superlinearly in the SPREAD coefficient, as Figures 8d and 8e indicate. In contrast to MINDIST, too large values will annihilate the separability, which becomes evident for the cluster triple $4, 7, 8$ in Fig. 8e.

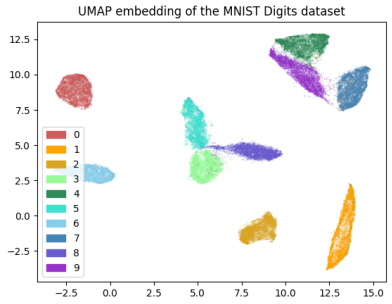
On the contrary, one should also avoid to choose SPREAD values smaller than 1, for the reason that they do not contract the points to the center as one might have expected, but drag them far away to the other side of the plot. After meeting in the middle (Fig. 8b), the clusters are lengthened along the vector that orthogonally maps from the plot center to the cluster (Fig. 8a). The pattern is somewhat fascinating and still allows to broadly distinguish some clusters by their original relative position to the center. However, keeping the recommended SPREAD $= 1.0$ seems to be a reasonable choice that may work for a huge fraction of the datasets to which UMAP is applied.



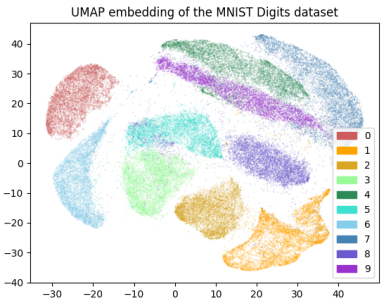
(a) SPREAD = 0.01



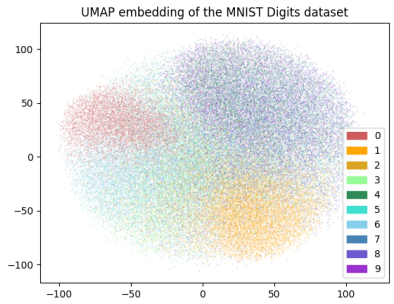
(b) SPREAD = 0.1



(c) SPREAD = 1.0



(d) SPREAD = 10



(e) SPREAD = 100

Figure 8: Spreading the data along various scales for $N_{\text{Neighbors}} = 50$ and MINDIST $= 0$. Similar results also hold for MINDIST $= 0.01, 0.1$. While an inverse spread contracts all points to the origin, a large spread will bloat the clusters such that they will eventually overlap.

4 Conclusion

The experiment visualized the expressivity of the first principal components in PCA. Furthermore, it confirmed the adequacy of the default hyperparameter choice for UMAP. Over and above, it shed light unto the impact of the distance metric. So far, the experiments provide no motivation to barter another metric for the Euclidean metric. It could be interesting to look for real world cases where the Euclidean distance provides only mediocre embeddings and other distances out of the colourful bouquet of options may be preferred.