CSIC 5011 Final Project : MNIST data reduction and performance analysis

Inam Ulhaqazad {iuaa@connect.ust.hk}
Department of Electronic and Computer Engineering, HKUST

1. Introduction

Principal component analysis (PCA) is a widely used technique for dimensionality reduction. Most datasets have correlated variables, and PCA finds a set of uncorrelated orthogonal components that capture the most information from data. In this project, we use Horn's parallel to find the number of relevant PCs. Then a neural network (MLP) is trained on the original dataset and reduced dataset comprising of PCs. The performance comparison is done for both cases.

2. MNIST Dataset

Modified National Institute of Standards and Technology (MNIST) dataset consists of 28*28 grayscale images of handwritten digits (0-9), along with their corresponding labels. With a total of 70,000 images, 60,000 images are used for training the model and the remaining 10,000 are reserved for testing its performance.

Data Preprocessing

The 28*28 matrix is flattened to get a single array of length 784 for each image. The pixel values are normalized by dividing by 255. Additionally, the data is centered by subtracting the mean from each value.

3. PCA using Horn's Parallel Analysis

PCA is given by the top k vectors of the Singular value decomposition (SVD) of the data matrix. The choice of k is given by Horn's parallel analysis. For each data point, the sample features are randomly permuted for decorrelation, and singular values of random matrices are calculated. This process is repeated several times to get a set of singular values. The p-value for each eigenvalue is defined and only those eigenvalues are kept whose p-value is smaller than a threshold. This gives a set of top k principal components corresponding to the eigenvalues.

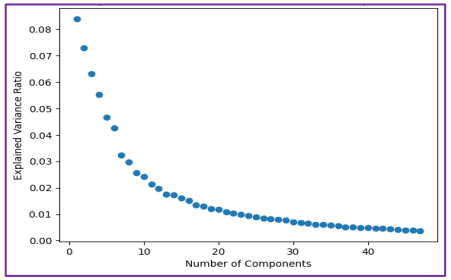
Equation $\sum_{i=1}^{k} \hat{\lambda}_i / \operatorname{tr}(\hat{\Sigma}_n) > q, \quad \text{e.g.} \quad q = 0.95$

4. Methodology

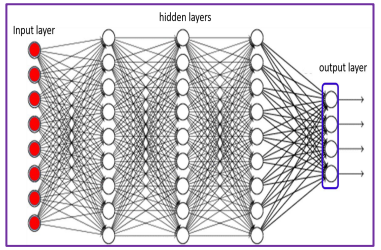
Deep learning (DL) architecture is used for the classification task. Initially, the model is trained on the original dataset having 784 features and performance is reported. Then PCA is used to do the feature reduction through Horn's parallel analysis. Subsequently, the data is represented by the top 47 principal components and the same DL model is used to report the performance metrics.

DL Architecture

Multi-layered perceptron (MLP) with 3 hidden layers and '*Relu*' activation function is used. The number of neurons in the first hidden layer is 512, followed by 128 and 64 in the subsequent layers. Each hidden layer is followed by a dropout layer with a dropout rate of 0.2 to avoid overfitting. The model is trained for 5 epochs.



Top 47 PCs for explained variance ratio



MLP with 3 hidden layers

5. Results

The top 47 PCs capture the data effectively as per Horn's analysis. These 47 PCs collectively explain more than 80% of the total variance in the dataset.

The classification accuracy of the trained model on the test set in the first case (784 features) is 97.7%. This model has 470k trainable parameters. With reduced features (47 PCs), the accuracy increases a little to around 98%, while the trainable parameters are reduced to 99k.

6. Analysis

A significant amount of information is retained by the top 47 PCs which still capture more than 80% of the information contained in dataset. The amount of variance explained by each PC also decreases with each additional component and the first component captures the most variance.

PCA helps to reduce the features by an order of 10 and the model parameters also reduce by the same factor. This reduces the model complexity and decreases training time. The accuracy increases which implies that the PCs capture the relevant information sufficient for drawing classification boundaries. It might be possible that the original data has some noise, and all the features are not important for drawing perfect boundaries.

7. Conclusion

This work demonstrates that PCA is an effective dimensionality reduction technique that maintains the classification accuracy comparable to using the original features. While significantly reducing the feature space, PCA effectively retains the crucial information necessary for accurate classification. The performance can be further increased by using deeper models like Inception, AlexNet, etc.

The 2D visualization of PCs does not yield significant boundaries implying that we need the order of 10 PCs to have sufficient information about our data. t-SNE embeddings seem to provide a good visualization of the MNIST data in the 2D space.

8. References

1. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 2011; 12:2825–2830
2. Abadi M, Barham P, Chen J, et al. TensorFlow: A system for large-scale machine learning
3. Yao Yuan, Topological and Geometric Data Reduction and Visualization Spring 2024, Lecture Notes

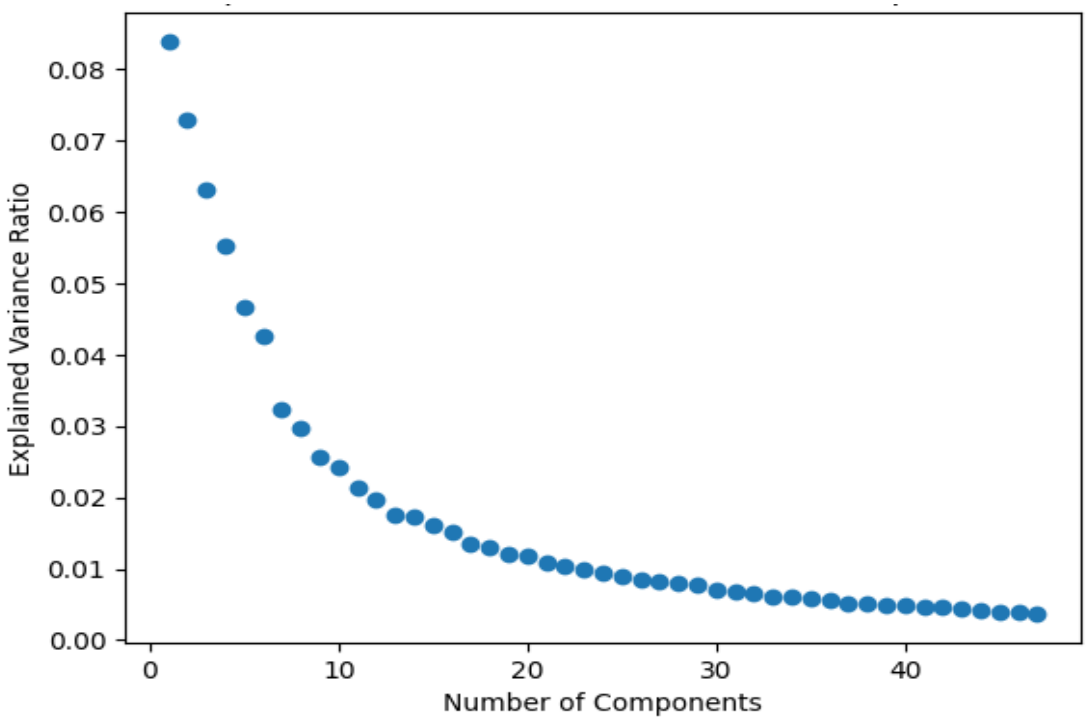


Fig. 1 Explained variance ratio

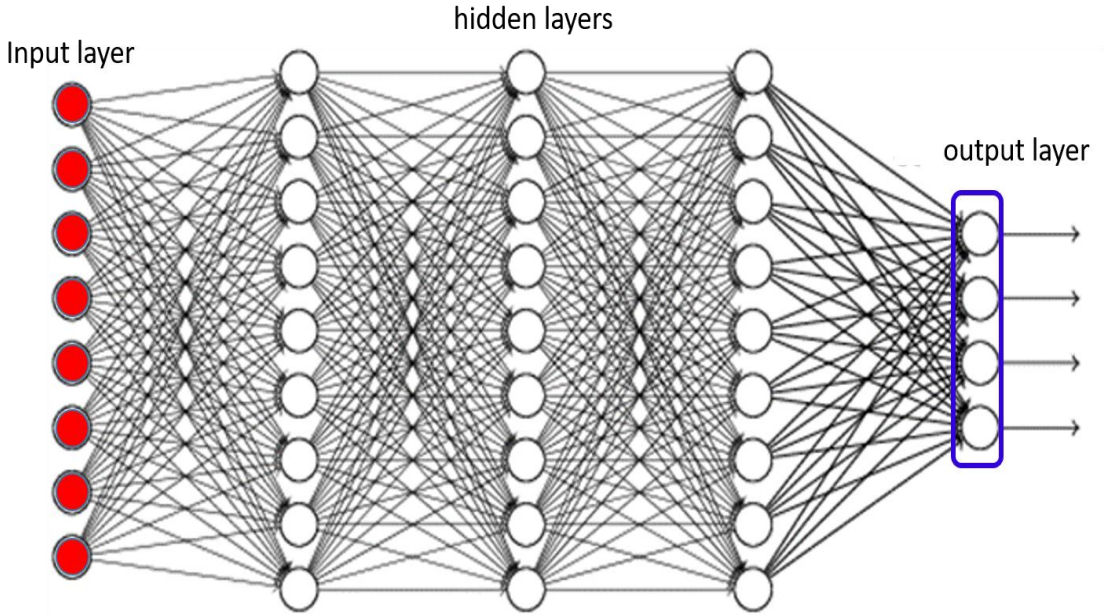


Fig. 2 MLP with 3 hidden layers

$$\sum_{i=1}^{k} \hat{\lambda}_i / \operatorname{tr}(\hat{\Sigma}_n) > q, \quad \text{e.g.} \quad q = 0.95$$

The equation for Horn's parallel analysis