INSI University

M1 Intelligence Artificielle et Data Science

Rapport de Recherche

Analyse Dimensionnelle du Jeu de Données MNIST

par Méthode des Composantes Principales

Auteurs :

RAHOLDINA FIARA Anjara Mihavana

RANDRIAMIARAMANANA Narivelo Yvan

ANDRIAMANAKOTO Anjara Tafita

RALALASON Rodéo Victorieux

Sous la direction de :

Dr. Dimby

Date de remise :

18 juin 2025

Année Universitaire 2023-2024

Master en Informatique

Table des matières

0.1 Introduction

L'analyse des données de grande dimension constitue un défi majeur dans le domaine de l'apprentissage automatique et de la vision par ordinateur. Le jeu de données MNIST, contenant 70 000 images de chiffres manuscrits représentés sous forme de matrices 28×28 pixels, illustre parfaitement cette problématique avec ses 784 dimensions initiales par observation.

La méthode des Composantes Principales (PCA) se révèle particulièrement adaptée pour aborder cette complexité dimensionnelle. Son principe fondamental repose sur la projection des données dans un sous-espace de plus faible dimension tout en maximisant la variance conservée. Cette approche permet non seulement de faciliter la visualisation des structures sous-jacentes aux données, mais également de réduire le temps de calcul pour les algorithmes d'apprentissage ultérieurs.

L'objectif principal de cette étude consiste à explorer systématiquement l'application de la PCA au jeu de données MNIST. Nous nous attacherons particulièrement à comprendre comment les différentes composantes principales capturent les variations essentielles entre les chiffres manuscrits et à évaluer la qualité des représentations dimensionnellement réduites.

0.2 Fondements Théoriques de l'Analyse en Composantes Principales

La PCA s'appuie sur des concepts mathématiques fondamentaux issus de l'algèbre linéaire et des statistiques multivariées. Soit une matrice de données centrées $X \in \mathbb{R}^{n \times p}$ où n représente le nombre d'observations et p le nombre de variables initiales (dans notre cas $p = 784$).

La première étape consiste au calcul de la matrice de covariance :

$$\Sigma = \frac{1}{n} X^T X \tag{1}$$

La diagonalisation de cette matrice symétrique semi-définie positive permet d'obtenir la décomposition spectrale :

$$\Sigma = W \Lambda W^T \tag{2}$$

où Λ est une matrice diagonale contenant les valeurs propres $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ et W est une matrice orthogonale dont les colonnes correspondent aux vecteurs propres

associés.

Les composantes principales s'obtiennent par projection :

$$Z = XW \tag{3}$$

Le choix du nombre k de composantes à retenir repose généralement sur le critère de la variance expliquée cumulée :

$$V(k) = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{p} \lambda_i} \tag{4}$$

0.3 Méthodologie Appliquée au Jeu de Données MNIST

0.3.1 Prétraitement des données

Avant application de la PCA, les données subissent une normalisation cruciale. Chaque pixel, initialement codé sur une échelle [0, 255], est transformé selon :

$$x'_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \tag{5}$$

où μ_j et σ_j représentent respectivement la moyenne et l'écart-type du j-ème pixel sur l'ensemble du jeu de données. Cette étape garantit que toutes les variables contribuent équitablement à l'analyse.

0.3.2 Implémentation pratique

L'implémentation utilise conjointement NumPy pour les calculs matriciels fondamentaux et scikit-learn pour une version optimisée :

```
# Version manuelle (p dagogique)
import numpy as np
cov_matrix = np.cov(X_std, rowvar=False)
eigenvals, eigenvecs = np.linalg.eigh(cov_matrix)
sorted_idx = np.argsort(eigenvals)[::-1]
components = eigenvecs[:, sorted_idx[:3]]
X_pca = X_std @ components

# Version scikit-learn (optimale)
from sklearn.decomposition import PCA
pca = PCA(n_components=3)
X_pca = pca.fit_transform(X_std)
```

Listing 1 – Implémentation de la PCA

0.3.3 Analyse de la variance

Le calcul systématique des ratios de variance expliquée permet de déterminer le nombre optimal de composantes à conserver pour différentes applications.

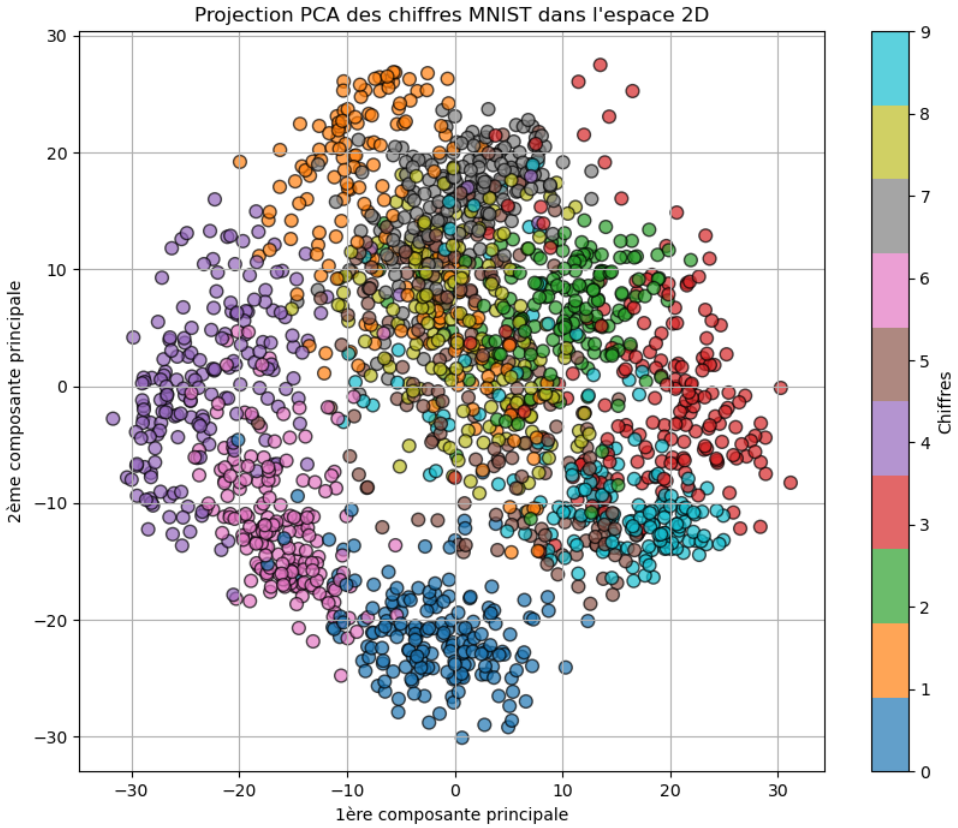

FIGURE 1 – Graphique de variance expliquée cumulée

0.4 Résultats et Analyse

0.4.1 Visualisation des composantes principales

La projection sur les deux premières composantes révèle une séparation partielle des classes. Les chiffres 0 et 1 apparaissent clairement isolés, tandis que les groupes 3-5-8 présentent des zones de chevauchement significatives.

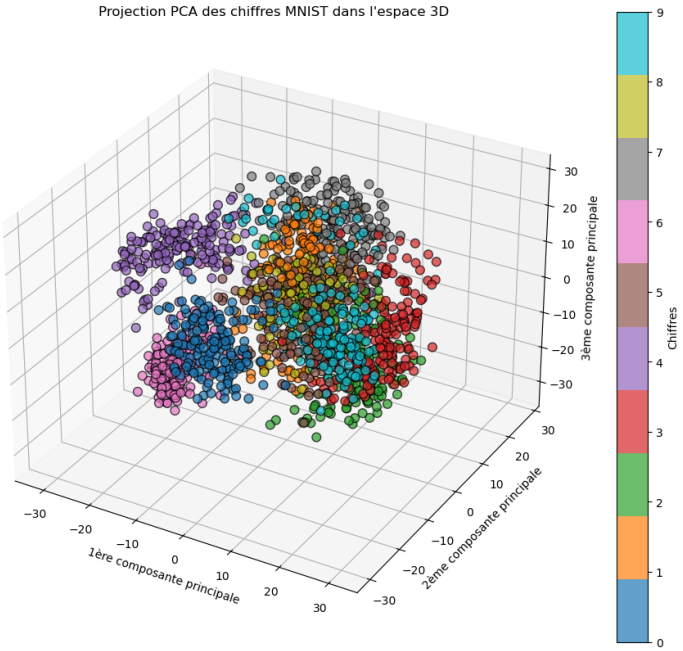

FIGURE 2 – Projection 2D des données MNIST sur les deux premières CP

L'analyse tridimensionnelle apporte un éclairage complémentaire en mettant en évidence des structures non visibles en 2D. La troisième composante principale semble particulièrement sensible aux courbures caractéristiques des chiffres 2 et 6.

0.4.2 Variance expliquée

Comme le montre la Figure 1, il faut conserver environ 200 composantes pour atteindre 95% de la variance totale. Ce résultat souligne la complexité intrinsèque du jeu de données MNIST.

0.5 Discussion Critique

Les résultats obtenus démontrent l'efficacité de la PCA pour la visualisation exploratoire des données MNIST. Cependant, plusieurs limitations méritent d'être soulignées.

La nature linéaire de la transformation PCA apparaît inadaptée pour capturer complètement les relations non linéaires entre certains chiffres morphologiquement proches. Par

ailleurs, l'interprétation physique des composantes au-delà de la troisième reste délicate.

Des méthodes alternatives comme t-SNE ou UMAP pourraient pallier certaines de ces limitations au prix d'une plus grande complexité algorithmique et d'une perte d'interprétabilité.

0.6 Conclusion et Perspectives

Cette étude approfondie a permis de mettre en lumière les apports et limites de la PCA appliquée au jeu de données MNIST. Les résultats obtenus ouvrent plusieurs pistes de recherche prometteuses.

L'extension aux variantes colorées de MNIST (comme Fashion-MNIST ou Color-MNIST) ou l'intégration avec des techniques de deep learning constituent des prolongements naturels de ce travail. Une autre voie intéressante consisterait à combiner la PCA avec des méthodes de classification pour évaluer l'impact de la réduction dimensionnelle sur les performances prédictives.

Annexes

— Code complet disponible sur https://github.com/Mihavana/mini-projet-mnist$_pca$

— Jeu de données MNIST original : `http://yann.lecun.com/exdb/mnist/`

— Scripts supplémentaires d'analyse et de visualisation

Références Bibliographiques

1. Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer
2. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer
3. Abdi, H., & Williams, L. J. (2010). Principal component analysis. *Wiley Interdisciplinary Reviews : Computational Statistics*, 2(4), 433-459.
4. Géron, A. (2019). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. O'Reilly.
5. Shlens, J. (2014). A tutorial on principal component analysis. arXiv preprint arXiv :1404.1100.