Réduction de dimension pour l'analyse et la visualisation de grands jeux de données multidimensionnels

Étude théorique et pratique

Rapport réalisé par :
Mohamed Issam El Khayati

Encadré par :
Khalide Jbilou

Année universitaire : 2025-2026

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Au cours des dernières décennies, l'évolution rapide des technologies de l'information, des capteurs numériques et des systèmes informatiques a conduit à une augmentation exponentielle de la quantité de données générées et stockées. Dans de nombreux domaines scientifiques et industriels tels que la vision par ordinateur, la reconnaissance de formes, le traitement du signal, la bio-informatique, la finance ou encore l'apprentissage automatique, les données manipulées sont de plus en plus volumineuses et caractérisées par un grand nombre de variables. Ces jeux de données multidimensionnels, bien que riches en information, soulèvent de nombreux défis en matière d'analyse, de modélisation, de visualisation et de complexité algorithmique.

L'un des principaux obstacles liés à l'exploitation de données de grande dimension est le phénomène connu sous le nom de *malédiction de la dimension* (*curse of dimensionality*). Lorsque la dimension de l'espace des caractéristiques augmente, le volume de cet espace croît de manière exponentielle, rendant les données de plus en plus clairsemées. Dans ce contexte, les notions classiques de distance et de similarité deviennent moins pertinentes, ce qui affecte directement les performances des algorithmes de classification, de regroupement et de reconnaissance de formes. Par ailleurs, le coût computationnel des algorithmes augmente considérablement, ce qui peut rendre leur utilisation impraticable pour des applications réelles à grande échelle.

Un autre problème majeur réside dans la difficulté d'interprétation et de visualisation des données multidimensionnelles. L'être humain est naturellement limité à la perception de données en deux ou trois dimensions, ce qui rend impossible la visualisation directe de données de grande dimension. Cette limitation complique l'exploration des données, la détection de structures cachées, ainsi que l'analyse qualitative des résultats fournis par les modèles d'apprentissage automatique.

Face à ces contraintes, la réduction de dimension s'impose comme une étape clé du processus d'analyse de données. Elle consiste à transformer les données initiales, souvent de grande dimension, vers un espace de dimension plus faible, tout en conservant autant que possible l'information pertinente. Cette transformation permet de réduire la redondance entre les variables, d'atténuer l'impact du bruit et de simplifier les modèles. En pratique, les techniques de réduction de dimension contribuent à améliorer les performances des algorithmes de classification et de reconnaissance, à accélérer les temps de calcul et à faciliter la visualisation et l'interprétation des données.

Les méthodes de réduction de dimension peuvent être classées en deux grandes catégories : les méthodes linéaires et les méthodes non linéaires. Les méthodes linéaires, telles que l'Analyse en Composantes Principales (PCA) et l'Analyse Discriminante Linéaire (LDA), reposent sur des transformations linéaires de l'espace des données. Elles sont largement utilisées en raison de leur simplicité conceptuelle, de leur robustesse et de leur efficacité computationnelle. Toutefois, ces approches supposent que la structure sous-jacente des données peut être représentée de manière adéquate dans un sous-espace linéaire, ce qui

peut constituer une limitation lorsque les données présentent des relations complexes et non linéaires.

Afin de surmonter ces limitations, des méthodes de réduction de dimension non linéaires ont été développées. Des techniques telles que Isomap et Laplacian Eigenmaps cherchent à préserver la géométrie intrinsèque des données en considérant qu'elles résident sur une variété de faible dimension immergée dans un espace de grande dimension. Ces approches permettent de capturer des structures locales ou globales complexes et sont particulièrement adaptées à des données issues de phénomènes non linéaires, comme les images, les signaux ou les formes.

L'objectif principal de ce travail est de proposer une étude théorique et pratique approfondie des principales méthodes de réduction de dimension, en mettant l'accent sur la comparaison entre les approches linéaires et non linéaires. Cette étude inclut l'analyse des principes mathématiques sous-jacents, l'implémentation des algorithmes en langage Python, ainsi que leur application à des problèmes concrets tels que la classification d'images, la reconnaissance de formes et le traitement du signal. Une attention particulière sera accordée à l'évaluation des performances, à la qualité des projections obtenues et à l'impact de la réduction de dimension sur les tâches d'apprentissage automatique.

Ce rapport est structuré comme suit. Après une présentation des concepts fondamentaux liés à la réduction de dimension, les méthodes linéaires et non linéaires seront étudiées de manière détaillée. Une comparaison théorique et expérimentale de ces méthodes sera ensuite réalisée afin de mettre en évidence leurs avantages et leurs limites respectives. Enfin, plusieurs applications pratiques seront présentées, suivies d'une discussion générale permettant de dégager des perspectives et des axes de recherche futurs dans le domaine de la réduction de dimension.

Chapitre 1

Concepts fondamentaux

1.1 Données multidimensionnelles

1.1.1 Définition et exemples

Les données multidimensionnelles désignent des ensembles de données dans lesquels chaque observation est décrite par un grand nombre de variables, également appelées caractéristiques ou attributs. Chaque observation peut ainsi être représentée comme un point ou un vecteur dans un espace de dimension élevée. Mathématiquement, une donnée multidimensionnelle est modélisée par un vecteur de la forme :

$$\mathbf{x} = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$$

où d représente la dimension de l'espace des caractéristiques, c'est-à-dire le nombre de variables décrivant chaque observation.

Lorsque d est élevé, on parle de données de grande dimension. Ce type de données est devenu omniprésent avec le développement des systèmes numériques modernes et des technologies de collecte automatique de données. Les données multidimensionnelles permettent de capturer une grande richesse d'information, mais leur complexité rend leur analyse plus difficile.

On rencontre ce type de données dans de nombreux domaines applicatifs, notamment :

— **Images** : une image numérique est constituée de pixels, chacun représentant une intensité lumineuse ou une valeur de couleur. Une image en niveaux de gris de taille $m \times n$ peut être représentée par un vecteur de dimension $m \times n$. Par exemple, une image de taille 28×28 pixels, comme celles du jeu de données MNIST, est représentée par un vecteur de dimension 784. Pour les images en couleur, la dimension est encore plus élevée en raison des différents canaux (RGB).

— **Signaux** : les signaux audio, radar ou biomédicaux (ECG, EEG, EMG) sont généralement échantillonnés dans le temps. Un signal enregistré sur une longue durée avec une fréquence d'échantillonnage élevée conduit à des vecteurs de grande dimension. Dans certains cas, des représentations fréquentielles (transformée de Fourier, ondelettes) augmentent encore la dimension des données.

— **Données textuelles** : dans le traitement automatique du langage naturel, les textes sont souvent représentés par des vecteurs de fréquences de mots (Bag-of-Words), des pondérations TF-IDF ou des embeddings distribués. Ces représentations peuvent atteindre plusieurs milliers, voire dizaines de milliers de dimensions.

Ces exemples illustrent que la haute dimension est une caractéristique intrinsèque de nombreux problèmes réels en science des données et en apprentissage automatique.

1.1.2 Problèmes liés à la haute dimension

La manipulation et l'analyse de données de grande dimension soulèvent de nombreuses difficultés, regroupées sous le terme de *malédiction de la dimension*. Ce concept met en évidence les effets négatifs de l'augmentation du nombre de dimensions sur les performances des algorithmes et sur la qualité des résultats obtenus.

Parmi les principaux problèmes liés à la haute dimension, on peut citer :

- **Augmentation du coût de calcul** : lorsque la dimension des données augmente, les besoins en mémoire et le temps de calcul des algorithmes croissent fortement. De nombreux algorithmes d'apprentissage ont une complexité qui dépend directement du nombre de dimensions, ce qui limite leur applicabilité à des données de très grande dimension.
- **Perte de pertinence des distances** : dans les espaces de grande dimension, les distances entre les points tendent à se concentrer, c'est-à-dire que la différence entre la distance minimale et la distance maximale devient de plus en plus faible. Ce phénomène rend les mesures de similarité moins discriminantes, ce qui affecte les algorithmes basés sur la distance, tels que les k-plus proches voisins (k-NN) ou les méthodes de clustering.
- **Sur-apprentissage** : un nombre élevé de caractéristiques augmente le risque de sur-apprentissage, en particulier lorsque le nombre d'exemples disponibles est limité. Le modèle peut alors s'adapter excessivement aux données d'entraînement et généraliser difficilement à de nouvelles données.
- **Difficulté de visualisation et d'interprétation** : il est impossible de visualiser directement des données au-delà de trois dimensions. Cette limitation complique l'analyse exploratoire des données et la compréhension des structures sous-jacentes.

Ces différentes difficultés montrent la nécessité de techniques permettant de réduire la complexité des données tout en conservant leur information essentielle, ce qui motive l'utilisation de la réduction de dimension.

1.2 Réduction de dimension

1.2.1 Définition

La réduction de dimension est une technique fondamentale en analyse de données qui vise à représenter des données de grande dimension dans un espace de dimension plus faible. L'objectif est de conserver autant que possible l'information pertinente tout en éliminant les redondances et le bruit.

Formellement, la réduction de dimension consiste à déterminer une fonction de projection :

$$f : \mathbb{R}^d \to \mathbb{R}^k \quad \text{avec } k \ll d$$

où d est la dimension initiale des données et k la dimension réduite. La fonction f peut être linéaire ou non linéaire selon la méthode employée.

Selon les cas, la structure à préserver peut correspondre à la variance globale des données, aux distances entre les points, aux relations de voisinage ou encore à l'information discriminante entre différentes classes.

1.2.2 Objectifs de la réduction de dimension

La réduction de dimension répond à plusieurs objectifs essentiels dans l'analyse et le traitement des données :

- **Visualisation** : projeter les données dans un espace bidimensionnel ou tridimensionnel afin de permettre leur visualisation. Cela facilite l'exploration des données, la détection de structures, de regroupements ou d'anomalies.
- **Compression de données** : représenter les données de manière plus compacte, en réduisant leur taille tout en conservant une approximation fidèle de l'information originale. Cet aspect est particulièrement important pour le stockage et la transmission des données.
- **Amélioration des performances en apprentissage automatique** : en supprimant les caractéristiques redondantes ou bruitées, la réduction de dimension peut améliorer la précision, la robustesse et la capacité de généralisation des algorithmes de classification et de reconnaissance.

Ainsi, la réduction de dimension constitue souvent une étape de prétraitement indispensable dans les pipelines d'apprentissage automatique modernes.

1.2.3 Réduction de dimension vs sélection de caractéristiques

Il est important de distinguer la réduction de dimension de la sélection de caractéristiques, deux approches souvent confondues mais conceptuellement différentes.

La sélection de caractéristiques consiste à choisir un sous-ensemble des variables originales jugées les plus pertinentes pour une tâche donnée. Les caractéristiques sélectionnées restent inchangées, ce qui facilite l'interprétation des résultats.

En revanche, la réduction de dimension crée de nouvelles variables, généralement obtenues comme des combinaisons linéaires ou non linéaires des caractéristiques initiales. Ces nouvelles variables, appelées composantes ou coordonnées latentes, permettent d'obtenir une représentation plus compacte et souvent plus informative des données.

1.3 Méthodes linéaires vs non linéaires

1.3.1 Principes généraux

Les méthodes de réduction de dimension peuvent être regroupées en deux grandes familles selon la nature de la transformation appliquée aux données :

- **Méthodes linéaires** : elles supposent que les données peuvent être représentées efficacement dans un sous-espace linéaire de dimension réduite. Les projections sont obtenues par des transformations linéaires, souvent basées sur des outils algébriques tels que les valeurs propres et les vecteurs propres.
- **Méthodes non linéaires** : elles cherchent à capturer la structure intrinsèque non linéaire des données, en supposant que celles-ci résident sur une variété de faible dimension immergée dans un espace de grande dimension. Ces méthodes visent à préserver les relations locales ou globales entre les points.

1.3.2 Cas d'utilisation typiques

Les méthodes linéaires, telles que l'Analyse en Composantes Principales (PCA) et l'Analyse Discriminante Linéaire (LDA), sont largement utilisées en pratique en raison de

leur simplicité, de leur efficacité computationnelle et de leur bonne interprétabilité. Elles sont particulièrement adaptées lorsque les relations entre les variables sont approximativement linéaires et lorsque le nombre de dimensions est très élevé.

Les méthodes non linéaires, comme Isomap et Laplacian Eigenmaps, sont plus appropriées lorsque les données présentent des structures complexes et non linéaires, telles que des formes courbes ou des variétés. Elles sont souvent utilisées pour la visualisation, l'exploration de données complexes et l'analyse de structures latentes, notamment dans les domaines de la vision par ordinateur et de la reconnaissance de formes.

Chapitre 2

Méthodes linéaires de réduction de dimension

Les méthodes linéaires de réduction de dimension reposent sur l'hypothèse que les données peuvent être représentées de manière satisfaisante dans un sous-espace linéaire de dimension plus faible que l'espace original. Ces méthodes sont largement utilisées en pratique en raison de leur simplicité, de leur interprétabilité et de leur efficacité computationnelle. Parmi les techniques les plus connues, on trouve l'Analyse en Composantes Principales (PCA) et l'Analyse Discriminante Linéaire (LDA), qui diffèrent principalement par leur caractère supervisé ou non supervisé et par leurs objectifs respectifs.

2.1 Analyse en composantes principales (PCA)

2.1.1 Cadre mathématique

Soit un ensemble de données constitué de n observations $\{x_1, x_2, \ldots, x_n\} \subset \mathbb{R}^d$, où chaque observation x_i représente un point dans un espace de dimension d. On suppose que les données sont centrées, c'est-à-dire que la moyenne empirique est nulle :

$$\frac{1}{n}\sum_{i=1}^{n} x_i = 0.$$

On définit alors la matrice de covariance empirique par :

$$\Sigma = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^\top \in \mathbb{R}^{d\times d}.$$

La matrice Σ est symétrique et définie positive, ce qui garantit l'existence d'une base orthonormée de vecteurs propres.

L'objectif de la PCA est de projeter les données dans un sous-espace de dimension $k \ll d$ tout en conservant un maximum d'information, mesurée par la variance totale [1].

2.1.2 Formulation variationnelle

La première composante principale est définie comme la direction unitaire $u_1 \in \mathbb{R}^d$ qui maximise la variance des données projetées :

$$u_1 = \arg\max_{\|u\|=1} \frac{1}{n}\sum_{i=1}^{n} (u^\top x_i)^2.$$

Ce problème peut s'écrire sous forme matricielle :

$$\max_{\|u\|=1} u^\top \Sigma u.$$

En introduisant un multiplicateur de Lagrange λ pour la contrainte $\|u\|^2 = 1$, on obtient le lagrangien :

$$\mathcal{L}(u, \lambda) = u^\top \Sigma u - \lambda(u^\top u - 1).$$

La condition d'optimalité donne :

$$\Sigma u = \lambda u.$$

Ainsi, les directions principales sont les vecteurs propres de la matrice de covariance, et les valeurs propres associées représentent la variance expliquée par chaque composante.

2.1.3 Extension à plusieurs composantes

Les k premières composantes principales correspondent aux k vecteurs propres associés aux plus grandes valeurs propres de Σ :

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d \geq 0.$$

La projection d'un point $x \in \mathbb{R}^d$ sur le sous-espace principal de dimension k s'écrit :

$$z = U_k^\top x,$$

où $U_k = [u_1, u_2, \ldots, u_k]$ est la matrice formée des k premiers vecteurs propres.

2.1.4 Interprétation par minimisation de l'erreur de reconstruction

La PCA peut également être interprétée comme la solution du problème de minimisation de l'erreur quadratique de reconstruction :

$$\min_{\dim(V)=k} \sum_{i=1}^{n} \|x_i - P_V x_i\|^2,$$

où P_V désigne le projecteur orthogonal sur le sous-espace V.

Il est démontré que le sous-espace engendré par les k premiers vecteurs propres de Σ minimise cette erreur. Cette propriété confère à la PCA un caractère optimal parmi toutes les méthodes de projection linéaire de rang k.

2.1.5 Lien avec la décomposition en valeurs singulières

Soit $X \in \mathbb{R}^{n \times d}$ la matrice de données centrées. La décomposition en valeurs singulières (SVD) de X est donnée par :

$$X = U \Sigma_X V^\top,$$

où les colonnes de V sont les vecteurs propres de la matrice de covariance $\frac{1}{n} X^\top X$.

Ainsi, la PCA peut être efficacement calculée à l'aide de la SVD, ce qui permet de traiter des jeux de données de grande dimension sans former explicitement la matrice de covariance.

2.1.6 Propriétés et limites théoriques

La PCA présente plusieurs propriétés fondamentales :
— elle fournit une base orthonormée optimale pour la représentation linéaire des données ;
— elle maximise la variance projetée et minimise l'erreur de reconstruction ;
— elle est indépendante de toute information de classe.

Cependant, la PCA repose sur des hypothèses fortes de linéarité et de variance globale. Elle n'est pas adaptée à la modélisation de structures non linéaires complexes et peut être sensible à la présence de bruit ou de valeurs aberrantes.

Ces limitations motivent le recours à des méthodes non linéaires ou supervisées, telles que l'Isomap, les Laplacian Eigenmaps ou la LDA.

2.2 Analyse discriminante linéaire (LDA)

2.2.1 Cadre supervisé et notations

Soit un ensemble de données supervisées $\{(x_i, y_i)\}_{i=1}^n$, où $x_i \in \mathbb{R}^d$ représente une observation et $y_i \in \{1, \ldots, C\}$ son étiquette de classe. On note C le nombre total de classes et n_c le nombre d'échantillons appartenant à la classe c.

On définit la moyenne globale :

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i,$$

et la moyenne de chaque classe :

$$\mu_c = \frac{1}{n_c} \sum_{y_i=c} x_i.$$

L'objectif de la LDA est de trouver une projection linéaire permettant de maximiser la séparabilité entre les classes tout en minimisant la dispersion des données à l'intérieur de chaque classe[2].

2.2.2 Matrices de dispersion

La LDA repose sur la définition de deux matrices fondamentales :
— la matrice de dispersion intra-classe :

$$S_W = \sum_{c=1}^{C} \sum_{y_i=c} (x_i - \mu_c)(x_i - \mu_c)^\top;$$

— la matrice de dispersion inter-classes :

$$S_B = \sum_{c=1}^{C} n_c (\mu_c - \mu)(\mu_c - \mu)^\top.$$

La matrice S_W mesure la variabilité des données à l'intérieur des classes, tandis que S_B quantifie la séparation entre les centres de gravité des classes.

2.2.3 Critère de Fisher

La projection optimale est obtenue en maximisant le critère de Fisher :

$$J(w) = \frac{w^\top S_B w}{w^\top S_W w},$$

où $w \in \mathbb{R}^d$ est un vecteur de projection non nul.

Ce critère favorise les directions pour lesquelles la dispersion inter-classes est grande relativement à la dispersion intra-classe.

2.2.4 Résolution du problème d'optimisation

Le problème d'optimisation s'écrit :

$$\max_{w \neq 0} \frac{w^\top S_B w}{w^\top S_W w}.$$

Il se ramène à la résolution du problème aux valeurs propres généralisé :

$$S_B w = \lambda S_W w.$$

Les directions discriminantes correspondent aux vecteurs propres associés aux plus grandes valeurs propres λ. Le nombre maximal de directions discriminantes est limité par :

$$\text{rang}(S_B) \leq C - 1.$$

Ainsi, la LDA ne peut produire au maximum que $C - 1$ composantes, ce qui constitue une limitation intrinsèque de la méthode.

2.2.5 Projection et classification

Soit $W_k = [w_1, \ldots, w_k]$ la matrice formée des k premières directions discriminantes. La projection d'une observation x dans l'espace réduit est donnée par :

$$z = W_k^\top x.$$

Dans cet espace, la classification peut être réalisée à l'aide d'un classifieur simple tel que le plus proche voisin (KNN) ou un classifieur bayésien sous hypothèse gaussienne.

2.2.6 Interprétation probabiliste

Sous l'hypothèse que les données de chaque classe suivent une loi normale multivariée de même matrice de covariance, la LDA correspond au classifieur bayésien optimal. Cette interprétation probabiliste renforce la cohérence théorique de la méthode dans des contextes où ces hypothèses sont raisonnablement satisfaites.

2.2.7 Propriétés et limites

La LDA présente plusieurs avantages :
— elle exploite explicitement les informations de classe ;
— elle maximise la séparabilité entre classes ;
— elle fournit des projections interprétables.
Cependant, la méthode souffre de certaines limitations :
— elle est limitée à $C - 1$ dimensions ;
— elle suppose des distributions gaussiennes et des covariances similaires entre classes ;
— elle est sensible aux données bruitées et aux classes mal séparées.
Ces limites motivent l'utilisation de méthodes non linéaires ou hybrides lorsque les structures de données deviennent complexes.

2.3 Différences entre la PCA et la LDA

La PCA (Analyse en composantes principales) et la LDA (Analyse discriminante linéaire) sont deux méthodes linéaires de réduction de dimension largement utilisées, mais reposant sur des objectifs fondamentalement différents. Cette section met en évidence leurs principes, avantages et limites respectifs.

2.3.1 Différences conceptuelles

La différence principale entre la PCA et la LDA réside dans la nature de l'information utilisée pour construire la projection.

La PCA est une méthode **non supervisée** qui ne tient pas compte des étiquettes de classe. Elle cherche à maximiser la variance globale des données projetées, sans se soucier de la séparation entre classes. En conséquence, les directions retenues correspondent aux axes de plus grande dispersion des données.

À l'inverse, la LDA est une méthode **supervisée** qui exploite explicitement les informations de classe. Elle vise à maximiser la séparabilité inter-classes tout en minimisant la dispersion intra-classe, ce qui la rend particulièrement adaptée aux tâches de classification.

2.3.2 Avantages de la PCA

La PCA présente plusieurs avantages notables :
— simplicité conceptuelle et implémentation efficace ;
— faible coût computationnel, même pour des données de grande dimension ;
— robustesse relative au bruit grâce à la capture de la variance globale ;
— absence de dépendance aux labels, ce qui la rend applicable à des données non annotées.
Ces propriétés font de la PCA une méthode de référence pour l'exploration des données, la compression et la réduction de dimension préalable à d'autres algorithmes.

2.3.3 Limites de la PCA

Malgré ses atouts, la PCA présente certaines limitations :
— incapacité à exploiter les informations de classe ;
— séparation des classes non garantie dans l'espace projeté ;

— hypothèse implicite de linéarité ;
— sensibilité aux valeurs aberrantes.

Ainsi, une forte variance n'implique pas nécessairement une forte capacité discriminante.

2.3.4 Avantages de la LDA

La LDA se distingue par les avantages suivants :
— utilisation explicite des étiquettes de classe ;
— maximisation directe de la séparabilité entre classes ;
— projections plus interprétables pour la classification ;
— efficacité lorsque le nombre de classes est limité.

Ces caractéristiques expliquent les bonnes performances de la LDA dans des contextes supervisés bien structurés.

2.3.5 Limites de la LDA

La LDA souffre toutefois de contraintes importantes :
— nombre maximal de dimensions limité à $C - 1$;
— dépendance à des hypothèses statistiques fortes (normalité et covariances identiques) ;
— sensibilité au bruit et aux classes mal séparées ;
— performances dégradées lorsque les distributions sont complexes ou non linéaires.

Ces limites expliquent pourquoi la LDA peut être moins performante que la PCA dans certains jeux de données de grande dimension.

2.3.6 Synthèse comparative

En résumé, la PCA est particulièrement adaptée à la réduction de dimension et à l'analyse exploratoire lorsque les données ne sont pas annotées, tandis que la LDA est plus appropriée pour des tâches de classification supervisée visant une séparation explicite des classes.

Le choix entre ces deux méthodes dépend donc étroitement de la nature des données et de l'objectif expérimental visé.

Critère	**PCA**	**LDA**
Type de méthode	Non supervisée	Supervisée
Objectif principal	Maximiser la variance globale des données	Maximiser la séparabilité entre classes
Utilisation des labels	Non	Oui
Critère d'optimisation	Maximisation de $u^\top \Sigma u$	Maximisation de $\frac{w^\top S_B w}{w^\top S_W w}$
Nombre maximal de composantes	d	$C-1$
Capacité de séparation des classes	Non garantie	Élevée lorsque les classes sont bien définies
Robustesse au bruit	Bonne (variance globale)	Plus sensible au bruit intra-classe
Hypothèses statistiques	Faibles	Normalité et covariances similaires
Complexité computationnelle	Faible à modérée	Modérée
Domaines d'application	Exploration, compression, prétraitement	Classification supervisée

TABLE 2.1 – Comparaison entre la PCA et la LDA

Chapitre 3

Méthodes non linéaires de réduction de dimension

Les méthodes non linéaires de réduction de dimension ont été développées afin de dépasser les limitations des approches linéaires lorsque les données présentent des structures complexes. Dans de nombreux problèmes réels, les données de grande dimension sont supposées résider sur une variété non linéaire de faible dimension immergée dans un espace de grande dimension. Les méthodes non linéaires cherchent alors à découvrir cette structure intrinsèque en préservant certaines propriétés géométriques des données, telles que les distances géodésiques ou les relations de voisinage.

Parmi les méthodes non linéaires les plus connues, Isomap et Laplacian Eigenmaps occupent une place importante. Bien qu'elles reposent toutes deux sur des représentations en graphe, leurs objectifs et leurs principes diffèrent sensiblement.

3.1 Isomap (Isometric Mapping)

3.1.1 Motivation et cadre géométrique

Isomap est une méthode de réduction de dimension **non linéaire** fondée sur l'hypothèse que les données de grande dimension $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ sont échantillonnées à partir d'une variété lisse $\mathcal{M}$ de dimension intrinsèque $k \ll d$, immergée dans l'espace ambiant.

Contrairement à la PCA, qui suppose une structure linéaire globale, Isomap cherche à préserver la géométrie intrinsèque de la variété, mesurée par les **distances géodésiques** entre les points[3].

3.1.2 Construction du graphe de voisinage

La première étape consiste à construire un graphe non orienté $G = (V, E)$, où chaque sommet correspond à un point de données. Deux stratégies sont couramment utilisées :

— graphe des k plus proches voisins ;

— graphe ε-voisinage.

Une arête est ajoutée entre deux points voisins x_i et x_j, pondérée par leur distance euclidienne :

$$w_{ij} = \|x_i - x_j\|.$$

Ce graphe constitue une approximation discrète de la variété sous-jacente.

3.1.3 Approximation des distances géodésiques

La distance géodésique réelle $d_{\mathcal{M}}(x_i, x_j)$ sur la variété est inconnue. Isomap l'approxime par la plus courte distance sur le graphe :

$$\hat{d}_{ij} = \min_{\text{chemins } i \to j} \sum w_{pq}.$$

Ces distances sont calculées à l'aide d'algorithmes classiques de plus court chemin, tels que Dijkstra ou Floyd–Warshall.

On obtient ainsi une matrice de distances géodésiques estimées :

$$D_G = (\hat{d}_{ij})_{1 \leq i,j \leq n}.$$

3.1.4 Plongement métrique par MDS classique

Une fois les distances géodésiques estimées, Isomap cherche un plongement euclidien $\{z_1, \ldots, z_n\} \subset \mathbb{R}^k$ tel que :

$$\|z_i - z_j\| \approx \hat{d}_{ij}.$$

Ce problème est résolu par l'analyse multidimensionnelle classique (MDS), qui repose sur la double centration de la matrice des distances au carré :

$$B = -\frac{1}{2} H D_G^2 H, \quad H = I - \frac{1}{n} \mathbf{1}\mathbf{1}^\top.$$

La matrice B est une matrice de Gram, dont la décomposition spectrale donne :

$$B = V \Lambda V^\top.$$

Les coordonnées réduites sont alors obtenues par :

$$Z_k = V_k \Lambda_k^{1/2},$$

où Λ_k contient les k plus grandes valeurs propres positives.

3.1.5 Interprétation théorique

Isomap peut être interprété comme une extension non linéaire de la PCA, où la distance euclidienne est remplacée par une approximation de la distance géodésique sur la variété.

Lorsque la variété est convexe et suffisamment échantillonnée, il peut être montré que les distances géodésiques estimées convergent vers les distances réelles, garantissant ainsi la cohérence asymptotique de l'algorithme.

3.1.6 Propriétés et limites

Isomap possède plusieurs propriétés importantes :
— préservation de la structure géométrique globale ;
— capacité à dérouler des variétés non linéaires ;
— interprétation géométrique claire.

Cependant, la méthode présente également des limitations :
— forte dépendance au choix du nombre de voisins ;
— sensibilité au bruit et aux points aberrants ;
— coût computationnel élevé ($\mathcal{O}(n^3)$ dans le pire cas) ;
— difficulté à traiter de très grands jeux de données.

3.1.7 Lien avec les résultats expérimentaux

Les performances observées sur le jeu de données COIL-20 s'expliquent par la capacité d'Isomap à modéliser des trajectoires continues correspondant aux variations d'angle de vue. En revanche, sur MNIST, où les structures sont plus complexes et moins lisses, la méthode montre des limites en classification.

Ces observations confirment que l'efficacité d'Isomap dépend fortement de l'adéquation entre la géométrie des données et l'hypothèse de variété sous-jacente.

3.2 Laplacian Eigenmaps

3.2.1 Motivation et cadre théorique

Laplacian Eigenmaps est une méthode de réduction de dimension **non linéaire** fondée sur la théorie des graphes et l'approximation discrète de l'opérateur de Laplace–Beltrami sur une variété sous-jacente.

On suppose que les données $\{x_1, \ldots, x_n\} \subset \mathbb{R}^d$ sont échantillonnées à partir d'une variété lisse $\mathcal{M}$ de dimension intrinsèque $k \ll d$, et que la structure locale de cette variété contient l'information géométrique la plus pertinente[4].

3.2.2 Construction du graphe de similarité

La première étape consiste à construire un graphe pondéré $G = (V, E)$, où chaque sommet correspond à une observation. Deux points x_i et x_j sont connectés s'ils sont voisins selon un critère k-NN ou ε-voisinage.

Les poids des arêtes sont définis par une fonction de similarité :

$$w_{ij} = \begin{cases} \exp\left(-\dfrac{\|x_i - x_j\|^2}{t}\right), & \text{si } x_i \sim x_j, \\ 0, & \text{sinon,} \end{cases}$$

où $t > 0$ est un paramètre de largeur du noyau.

3.2.3 Matrices fondamentales du graphe

On définit la matrice d'adjacence pondérée $W = (w_{ij})$, ainsi que la matrice de degrés diagonale D :

$$D_{ii} = \sum_{j=1}^{n} w_{ij}.$$

Le Laplacien du graphe est alors défini par :

$$L = D - W.$$

Le Laplacien discret joue un rôle central dans la modélisation des variations locales de la fonction d'immersion sur le graphe.

3.2.4 Formulation variationnelle

L'objectif de Laplacian Eigenmaps est de trouver une représentation $\{z_1, \ldots, z_n\} \subset \mathbb{R}^k$ qui préserve la proximité locale :

$$\min \sum_{i,j} \|z_i - z_j\|^2 w_{ij}.$$

Cette fonctionnelle pénalise fortement les points proches dans l'espace original qui seraient éloignés dans l'espace réduit.

Sous la contrainte :

$$Z^\top D Z = I,$$

le problème se reformule en :

$$\min_Z \ \mathrm{Tr}(Z^\top L Z).$$

3.2.5 Problème spectral

La résolution conduit au problème aux valeurs propres généralisé :

$$Lz = \lambda D z.$$

Les vecteurs propres associés aux plus petites valeurs propres non nulles fournissent les coordonnées de l'espace réduit. La première valeur propre $\lambda_0 = 0$ correspond au vecteur constant et est ignorée.

Ainsi, les k vecteurs propres suivants définissent l'embedding :

$$Z = [z_1, \ldots, z_k].$$

3.2.6 Interprétation géométrique

Laplacian Eigenmaps peut être vu comme une approximation discrète de l'opérateur de Laplace–Beltrami sur la variété $\mathcal{M}$.

La méthode privilégie la conservation des relations locales plutôt que des distances globales, ce qui la rend particulièrement adaptée à l'analyse de structures locales complexes.

3.2.7 Propriétés et limites

Les principales propriétés de Laplacian Eigenmaps sont :

— préservation efficace de la structure locale ;
— forte robustesse aux déformations globales ;
— formulation mathématique élégante via le Laplacien.

Cependant, la méthode présente plusieurs limites :

— absence de garantie sur la préservation globale des distances ;
— dépendance au choix du graphe et des paramètres ;
— impossibilité d'extrapoler à de nouveaux points ;
— coût computationnel élevé pour de grands jeux de données.

3.3 Différences entre Isomap et Laplacian Eigenmaps

Isomap et Laplacian Eigenmaps sont deux méthodes de réduction de dimension **non linéaires** fondées sur la modélisation des données par un graphe de voisinage. Bien qu'elles partagent ce principe commun, leurs objectifs géométriques et leurs propriétés diffèrent sensiblement.

3.3.1 Différences conceptuelles

La différence fondamentale entre Isomap et Laplacian Eigenmaps réside dans le type de structure géométrique qu'elles cherchent à préserver.

Isomap vise à conserver la **géométrie globale** de la variété en approximant les distances géodésiques entre les points. Elle cherche à produire un plongement isométrique, où les distances dans l'espace réduit reflètent fidèlement les distances intrinsèques sur la variété.

À l'inverse, Laplacian Eigenmaps privilégie la **préservation de la structure locale**. La méthode ne cherche pas à conserver les distances globales, mais plutôt à maintenir la proximité entre points voisins, ce qui favorise la cohérence locale au détriment de la séparation globale des classes.

3.3.2 Avantages de Isomap

Isomap présente plusieurs avantages importants :
— capacité à dérouler des variétés non linéaires globales ;
— bonne préservation des distances intrinsèques ;
— interprétation géométrique claire ;
— efficacité sur des données organisées selon des trajectoires continues.

Ces propriétés expliquent son intérêt pour des jeux de données tels que COIL-20, où les variations sont régulières et fortement corrélées.

3.3.3 Limites de Isomap

Malgré ses atouts, Isomap présente certaines limitations :
— sensibilité élevée au choix du nombre de voisins ;
— forte dépendance à la connectivité du graphe ;
— coût computationnel élevé dû au calcul des plus courts chemins ;
— sensibilité au bruit et aux points aberrants.

Ces contraintes limitent son utilisation pour des jeux de données volumineux ou bruités.

3.3.4 Avantages de Laplacian Eigenmaps

Laplacian Eigenmaps se distingue par les avantages suivants :
— excellente préservation des relations locales ;
— robustesse face aux déformations globales ;
— formulation mathématique élégante via le Laplacien discret ;
— bonnes performances lorsque la structure locale est dominante.

Cette méthode est particulièrement efficace pour la visualisation et l'analyse exploratoire de données présentant de fortes corrélations locales.

3.3.5 Limites de Laplacian Eigenmaps

Laplacian Eigenmaps souffre néanmoins de plusieurs limitations :
— absence de garantie sur la structure globale ;
— séparation des classes non assurée ;
— dépendance au choix du graphe et du noyau ;
— impossibilité d'extrapoler facilement à de nouveaux points.
Ainsi, la méthode est moins adaptée aux tâches de classification directe.

3.3.6 Synthèse comparative

En résumé, Isomap est plus adaptée lorsque la structure globale de la variété est essentielle à l'analyse, tandis que Laplacian Eigenmaps est plus efficace pour préserver des relations locales fines.

Le choix entre ces deux méthodes dépend donc de la géométrie des données et de l'objectif expérimental, qu'il s'agisse de visualisation, d'exploration ou de classification.

Critère	**Isomap**	**Laplacian Eigenmaps**
Type de méthode	Non linéaire, basée sur la géométrie globale	Non linéaire, basée sur la structure locale
Structure préservée	Distances géodésiques globales	Relations locales de voisinage
Principe fondamental	Approximation des plus courts chemins sur le graphe	Minimisation d'une énergie locale basée sur le Laplacien
Outil mathématique principal	MDS classique sur distances géodésiques	Problème spectral du Laplacien du graphe
Sensibilité au choix des paramètres	Élevée (nombre de voisins, connectivité)	Élevée (voisinage, noyau, paramètre t)
Robustesse au bruit	Faible à modérée	Modérée pour les structures locales
Préservation de la séparation globale	Bonne si la variété est bien échantillonnée	Non garantie
Capacité de visualisation	Très bonne pour variétés continues	Très bonne pour structures locales complexes
Complexité computationnelle	Élevée (calcul des plus courts chemins)	Élevée (décomposition spectrale)
Adaptation à la classification	Limitée	Faible

Table 3.1 – Comparaison entre Isomap et Laplacian Eigenmaps

Chapitre 4

Comparaison des méthodes de réduction de dimension

Après avoir présenté les principales méthodes de réduction de dimension linéaires et non linéaires, il est essentiel de procéder à une comparaison approfondie afin de mieux comprendre leurs différences, leurs avantages et leurs limites. Cette comparaison permet de guider le choix de la méthode la plus adaptée en fonction de la nature des données, des objectifs de l'analyse et des contraintes computationnelles.

4.1 Critères de comparaison

Afin d'évaluer et de comparer efficacement les différentes méthodes de réduction de dimension étudiées, plusieurs critères sont pris en compte. Ces critères permettent d'analyser les performances des méthodes aussi bien d'un point de vue théorique que pratique, en tenant compte des besoins réels des applications.

4.1.1 Capacité de visualisation

La capacité de visualisation correspond à l'aptitude d'une méthode à projeter des données de grande dimension dans un espace de dimension 2 ou 3, tout en conservant une structure interprétable. Une bonne méthode de visualisation doit permettre de distinguer clairement les regroupements naturels, les relations entre les points ainsi que d'éventuelles séparations entre classes.

Les méthodes non linéaires, telles qu'Isomap et Laplacian Eigenmaps, sont généralement plus performantes pour la visualisation de données complexes présentant des structures non linéaires. En revanche, les méthodes linéaires comme la PCA offrent une représentation plus simple, plus stable et plus rapide à calculer, ce qui les rend très utilisées pour l'analyse exploratoire initiale.

4.1.2 Préservation de la structure des données

La préservation de la structure des données est un critère fondamental en réduction de dimension. Il s'agit de mesurer la capacité d'une méthode à conserver les relations importantes existant entre les données originales après projection dans un espace de plus faible dimension.

Selon la méthode utilisée, différentes structures peuvent être privilégiées. La PCA cherche principalement à conserver la variance globale des données, ce qui permet de préserver les directions les plus informatives. Isomap vise à préserver les distances géodésiques, assurant ainsi la conservation de la structure globale de la variété. Laplacian Eigenmaps, quant à elle, se concentre sur la préservation des relations locales entre les points voisins, ce qui est particulièrement utile pour l'analyse de structures fines.

4.1.3 Performance en classification

La réduction de dimension joue un rôle crucial dans les performances des algorithmes de classification. Une projection efficace doit améliorer la séparabilité des classes tout en réduisant le bruit et la redondance des caractéristiques.

La LDA, en tant que méthode supervisée, est spécifiquement conçue pour maximiser la séparation entre les classes, ce qui explique ses bonnes performances en classification lorsque les hypothèses du modèle sont respectées. Les méthodes non supervisées, telles que la PCA ou Isomap, ne tiennent pas compte des labels de classe et nécessitent souvent l'ajout d'un classifieur ou d'une étape d'optimisation supplémentaire pour obtenir de bonnes performances.

4.1.4 Complexité computationnelle

La complexité computationnelle est un critère déterminant, en particulier pour les grands jeux de données. Elle dépend à la fois du nombre d'échantillons, de la dimension initiale des données et de la nature de la méthode utilisée.

Les méthodes linéaires présentent généralement une complexité plus faible et une meilleure scalabilité. La PCA, par exemple, peut être implémentée efficacement à l'aide de techniques matricielles optimisées. En revanche, les méthodes non linéaires reposent souvent sur la construction et l'analyse de graphes, ainsi que sur le calcul de plus courts chemins ou de décompositions spectrales, ce qui entraîne un coût computationnel plus élevé.

4.1.5 Sensibilité au bruit

La sensibilité au bruit évalue la robustesse des méthodes face aux données aberrantes, aux mesures bruitées ou aux erreurs de capteurs. Une méthode robuste doit être capable de produire une projection stable malgré la présence de bruit.

Les méthodes linéaires, notamment la PCA, peuvent atténuer le bruit en éliminant les composantes de faible variance. En revanche, les méthodes basées sur les graphes, telles qu'Isomap et Laplacian Eigenmaps, sont souvent plus sensibles au bruit, car celui-ci peut affecter la construction du graphe de voisinage et dégrader la qualité de la projection.

4.2 Comparaison théorique

4.2.1 Méthodes linéaires vs non linéaires

D'un point de vue théorique, les méthodes linéaires supposent que les données peuvent être représentées efficacement dans un sous-espace linéaire de dimension réduite. Elles

sont simples à mettre en œuvre, rapides et facilement interprétables. Cependant, cette hypothèse limite leur capacité à représenter des structures complexes et non linéaires.

À l'inverse, les méthodes non linéaires sont capables de capturer la géométrie intrinsèque des données en tenant compte de la structure de variété. Elles sont donc mieux adaptées aux données réelles complexes, telles que les images ou les signaux. Néanmoins, ces méthodes sont plus coûteuses en calcul, plus sensibles aux paramètres et parfois plus difficiles à interpréter.

4.2.2 Méthodes supervisées vs non supervisées

Les méthodes supervisées, comme la LDA, exploitent explicitement les informations de classe pour orienter la projection dans un but précis, généralement la classification. Elles offrent d'excellentes performances lorsque les labels sont fiables et que les classes sont bien définies.

Les méthodes non supervisées, telles que la PCA, Isomap et Laplacian Eigenmaps, ne nécessitent pas de labels et sont donc particulièrement adaptées à l'exploration, à la visualisation et à l'analyse de données non annotées. Toutefois, elles ne garantissent pas une séparation optimale des classes dans l'espace projeté.

4.3 Comparaison expérimentale

4.3.1 Résultats quantitatifs

La comparaison expérimentale repose sur plusieurs indicateurs quantitatifs, tels que le taux de classification, l'erreur de reconstruction, la variance expliquée et le temps de calcul. Ces mesures permettent d'évaluer objectivement les performances des différentes méthodes dans des conditions contrôlées.

Les résultats obtenus montrent généralement que la LDA offre les meilleures performances en classification lorsque les classes sont bien séparées et que les hypothèses du modèle sont respectées. La PCA constitue un bon compromis entre performance et simplicité, tandis que les méthodes non linéaires fournissent des résultats particulièrement intéressants pour l'analyse exploratoire et la visualisation.

4.3.2 Analyse graphique

L'analyse graphique consiste à examiner visuellement les projections obtenues par chaque méthode dans un espace de dimension réduite. Ces visualisations permettent d'observer la séparation des classes, la formation de clusters et la préservation des structures géométriques.

Les méthodes non linéaires produisent souvent des visualisations plus expressives et plus fidèles à la structure intrinsèque des données, mettant en évidence des relations invisibles avec des méthodes linéaires. Toutefois, cette expressivité se fait au prix d'une complexité algorithmique et d'un temps de calcul plus élevés.

Chapitre 5

Application et résultats expérimentaux

5.1 Expérience 1 : Jeu de données MNIST

5.1.1 Jeu de données utilisé

Dans cette étude expérimentale, nous utilisons le jeu de données **MNIST**[5], qui constitue une référence incontournable dans le domaine de la reconnaissance de formes, de la vision par ordinateur et de l'apprentissage automatique. Ce jeu de données est largement utilisé pour évaluer les performances des algorithmes de classification et de réduction de dimension, en raison de sa simplicité apparente combinée à une grande variabilité intra-classe.

MNIST contient **70 000 images** de chiffres manuscrits allant de 0 à 9, réparties de la manière suivante :

— 60 000 images destinées à l'apprentissage,

— 10 000 images réservées à l'évaluation.

Chaque image est de taille 28×28 pixels et est représentée sous la forme d'un vecteur de dimension **784**. Cette dimension élevée rend l'analyse directe des données difficile, notamment en termes de visualisation et de coût computationnel, ce qui justifie pleinement l'utilisation de techniques de réduction de dimension.

Avant l'application des différentes méthodes étudiées, les données ont été **normalisées** à l'aide d'une standardisation (soustraction de la moyenne et division par l'écart-type). Cette étape est essentielle afin de garantir une meilleure stabilité numérique des algorithmes et d'éviter que certaines dimensions dominent artificiellement les autres lors des calculs de distance ou de variance.

5.1.2 Illustration du jeu de données

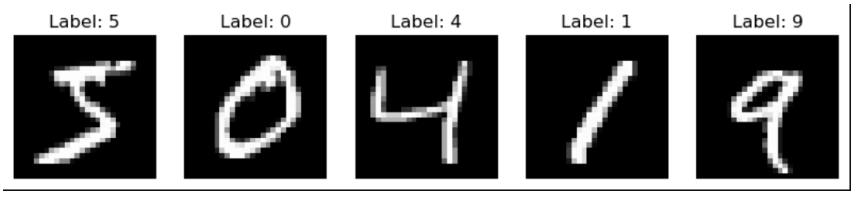

FIGURE 5.1 – Exemples d'images du jeu de données MNIST

La Figure 5.1 présente un ensemble d'exemples représentatifs issus du jeu de données MNIST. Chaque image correspond à un chiffre manuscrit, écrit par des personnes différentes, avec des styles d'écriture variés.

Interprétation : On observe une grande variabilité dans la forme, l'épaisseur des traits et l'orientation des chiffres. Certaines classes, comme les chiffres 1 et 7 ou encore 3 et 5, présentent des similitudes visuelles importantes. Cette variabilité intra-classe, combinée à des ressemblances inter-classes, rend la tâche de classification non triviale.

Cette figure sert de référence visuelle pour analyser l'impact des différentes méthodes de réduction de dimension sur la structuration des données et sur la séparation des classes dans les espaces projetés.

5.1.3 Visualisation par PCA

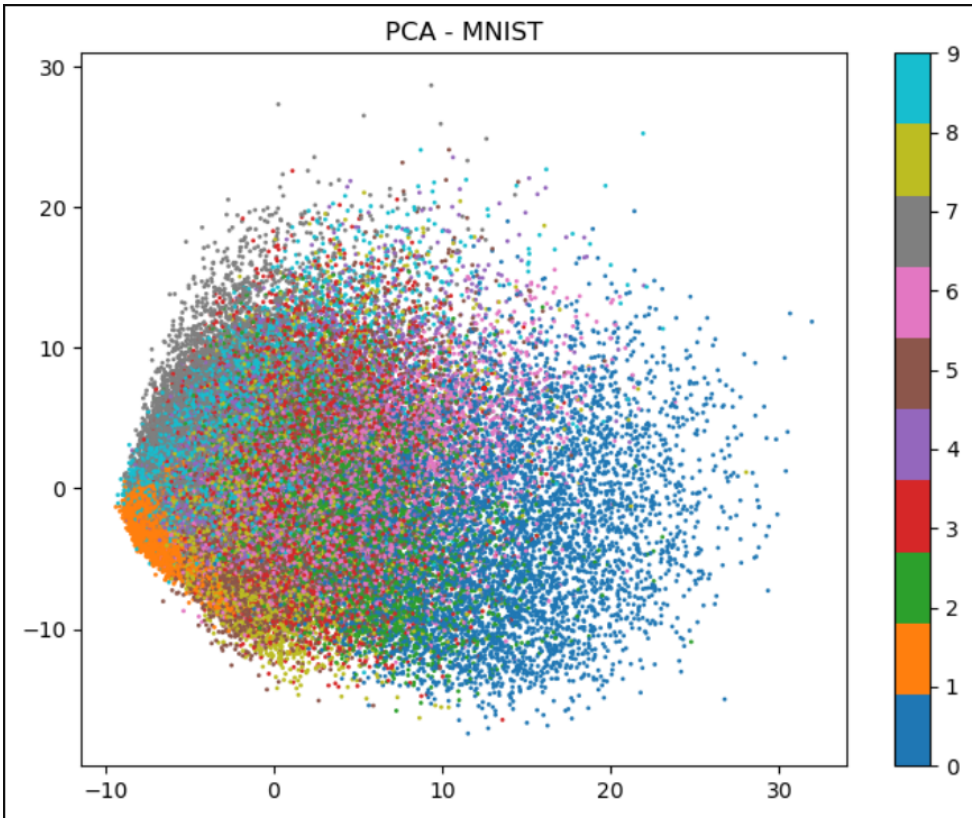

FIGURE 5.2 – Projection PCA en 2D du jeu de données MNIST

La Figure 5.2 montre la projection bidimensionnelle du jeu de données MNIST obtenue à l'aide de l'analyse en composantes principales (PCA).

Interprétation : La PCA vise à conserver les directions de plus grande variance des données, sans utiliser d'information sur les classes. On observe que certaines classes, telles que les chiffres 0, 1 et 2, forment des regroupements relativement cohérents. Toutefois, un chevauchement important subsiste entre plusieurs classes, notamment pour les chiffres ayant des formes similaires.

Malgré cette superposition, la structure globale des données est bien conservée. La précision obtenue avec un classifieur KNN appliqué aux données réduites à 50 dimensions atteint **96.02%**, ce qui montre que la PCA conserve une grande partie de l'information discriminante nécessaire à la classification.

Cette méthode constitue ainsi une approche efficace pour une première exploration des données et pour une réduction de dimension rapide et robuste.

5.1.4 Visualisation par LDA

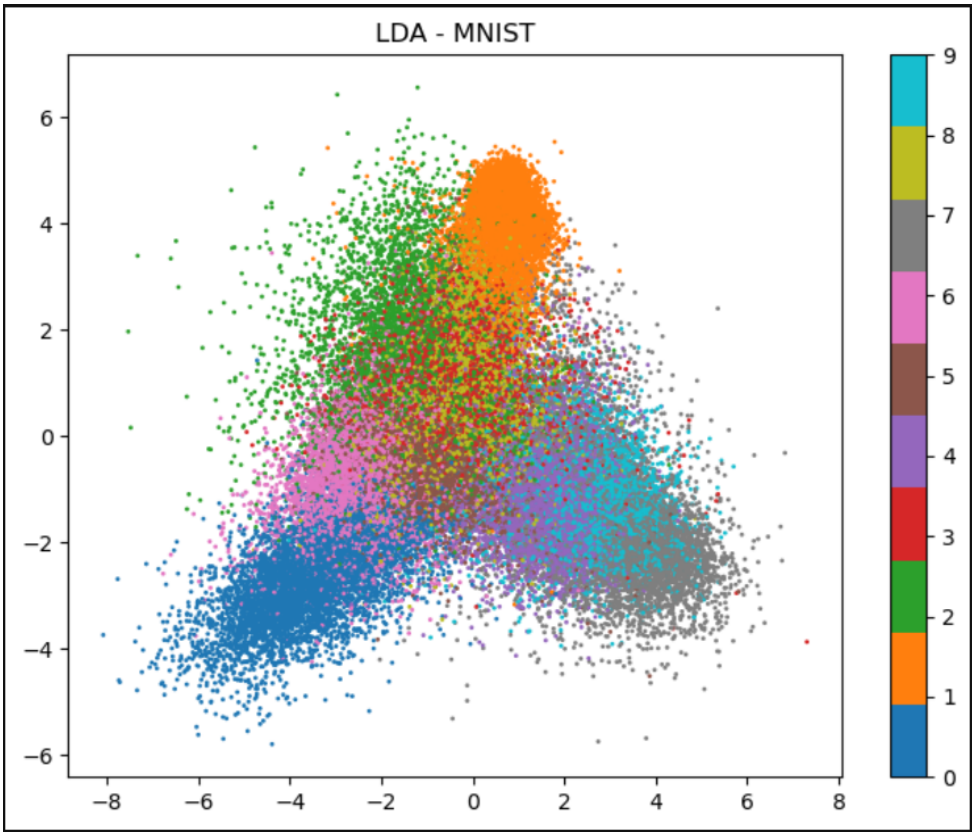

FIGURE 5.3 – Projection LDA en 2D du jeu de données MNIST

La Figure 5.3 présente la projection obtenue à l'aide de l'analyse discriminante linéaire (LDA).

Interprétation : Contrairement à la PCA, la LDA est une méthode supervisée qui exploite les informations de classe afin de maximiser la séparation entre celles-ci. On observe des clusters plus distincts pour plusieurs chiffres, notamment 0, 1, 6 et 7. Cependant, certaines classes, comme 4 et 9, restent partiellement superposées.

La précision obtenue par le classifieur KNN est de **92.05%**, légèrement inférieure à celle obtenue avec la PCA. Cette différence s'explique par la limitation intrinsèque de la LDA, qui ne peut produire au maximum que $n\_classes - 1$ composantes, soit 9 dans le cas de MNIST.

Malgré cette contrainte, la LDA offre une meilleure interprétabilité des axes projetés et illustre clairement l'intérêt de la supervision pour améliorer la séparabilité des classes.

5.1.5 Visualisation par Isomap

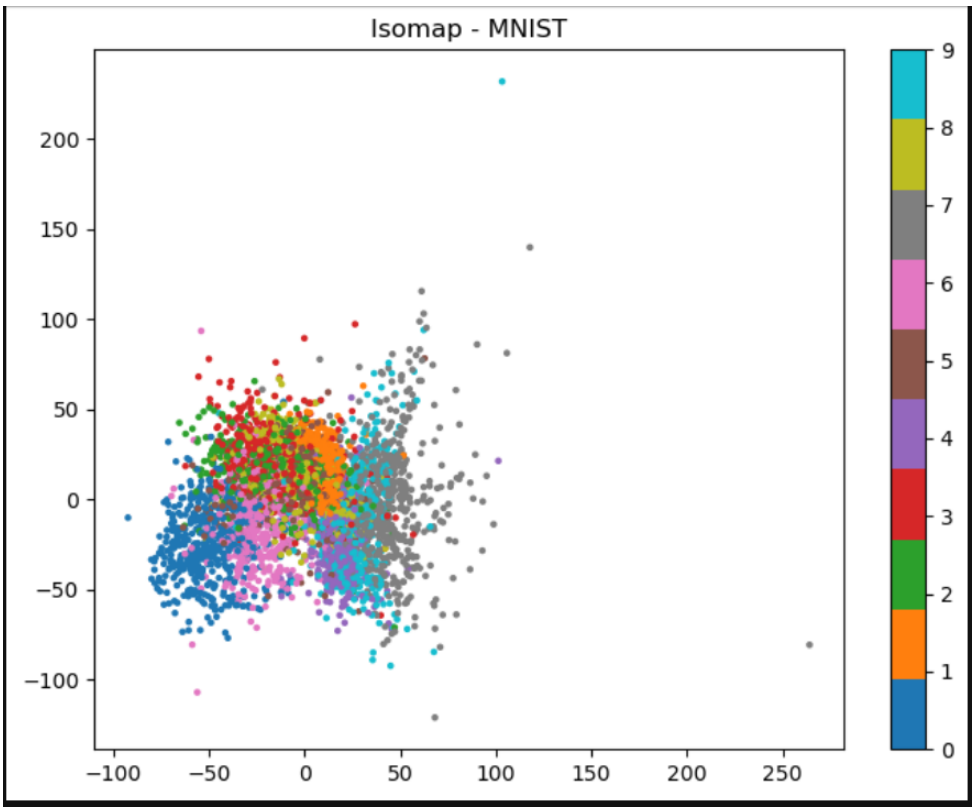

FIGURE 5.4 – Projection Isomap en 2D du jeu de données MNIST

La Figure 5.4 illustre la projection obtenue par la méthode Isomap à partir d'un sous-ensemble du jeu de données MNIST.

Interprétation : Isomap est une méthode non linéaire qui vise à préserver les distances géodésiques sur la variété sous-jacente des données. La projection met en évidence des structures courbes et des regroupements locaux correspondant aux relations complexes entre les images.

Certaines classes sont relativement bien séparées, mais des chevauchements persistent, notamment pour les chiffres visuellement proches comme 3, 5 et 6. La précision obtenue avec KNN est de **89.10%**, ce qui reste inférieur aux méthodes linéaires, en partie à cause du sous-échantillonnage et de la sensibilité aux paramètres du graphe.

Cette méthode est particulièrement intéressante pour révéler la structure intrinsèque des données, mais son coût computationnel élevé limite son application à de très grands ensembles.

5.1.6 Visualisation par Laplacian Eigenmaps

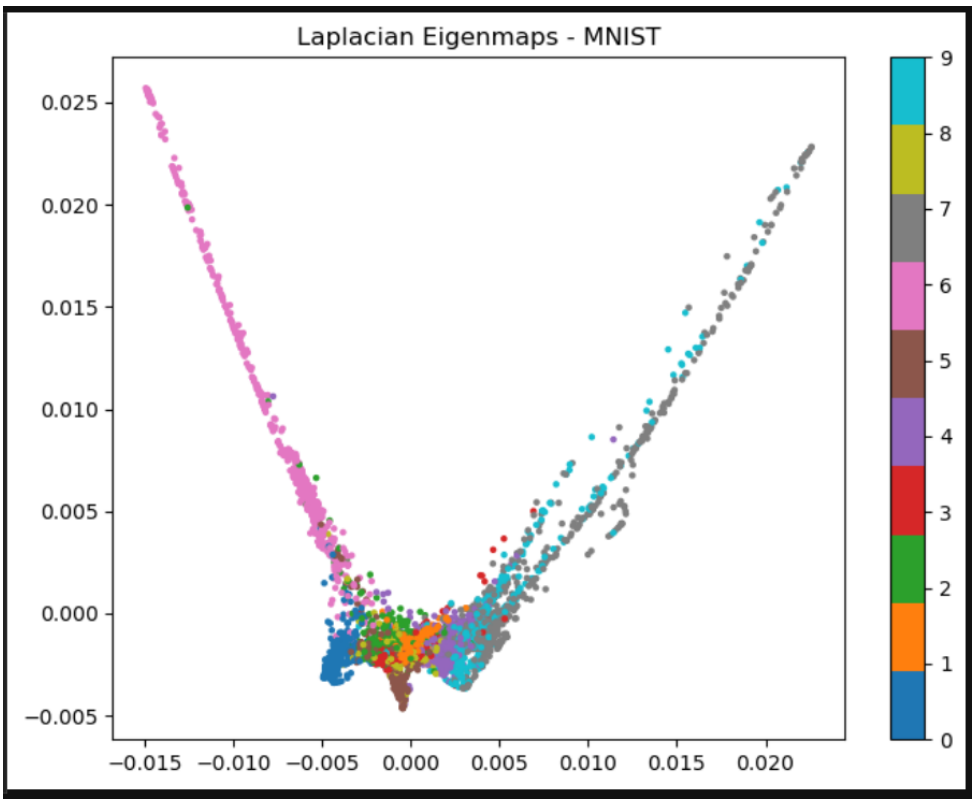

FIGURE 5.5 – Projection Laplacian Eigenmaps en 2D du jeu de données MNIST

La Figure 5.5 présente la projection obtenue à l'aide de la méthode Laplacian Eigenmaps.

Interprétation : Cette méthode met l'accent sur la préservation des relations locales entre les points voisins. On observe la formation de clusters locaux correspondant à certaines classes de chiffres, mais la séparation globale des classes est moins marquée.

Les chiffres 3, 5 et 8 apparaissent partiellement mélangés, ce qui reflète la priorité donnée à la structure locale plutôt qu'à la séparation globale. La précision KNN obtenue est de **88.60%**, confirmant que la préservation des voisinages locaux n'est pas toujours suffisante pour une classification optimale.

Méthode	**Taux de classification**
PCA + KNN	96.02%
LDA + KNN	92.05%
Isomap + KNN	89.10%
Laplacian Eigenmaps + KNN	88.60%

TABLE 5.1 – Performances de classification KNN selon la méthode de réduction de dimension

Ces résultats montrent que la PCA conserve une information globale suffisante pour obtenir d'excellentes performances de classification, tandis que les méthodes non linéaires sont davantage adaptées à la visualisation et à l'analyse exploratoire.

5.1.7 Discussion des résultats

L'analyse expérimentale menée sur le jeu de données MNIST met en évidence l'impact significatif du choix de la méthode de réduction de dimension sur la visualisation, la structure des données projetées et les performances de classification. Cette discussion vise à interpréter les résultats obtenus en tenant compte des propriétés théoriques de chaque méthode et de leurs comportements observés expérimentalement.

5.2 Expérience 2 : Jeu de données COIL-20

5.2.1 Jeu de données utilisé

La seconde expérience est menée sur le jeu de données **COIL-20** [6](Columbia Object Image Library), qui est couramment utilisé pour l'évaluation des méthodes de reconnaissance d'objets et de réduction de dimension.

COIL-20 est composé de **20 objets distincts**, chacun étant photographié sous différents angles de vue allant de 0° à 355°, avec un pas de 5°. Cela correspond à **72 images par objet**, soit un total de **1440 images**.

Les images sont initialement fournies en niveaux de gris. Dans cette étude, elles ont été redimensionnées à une taille de 32×32 pixels, puis vectorisées, conduisant à une représentation de dimension **1024** pour chaque image.

Avant l'application des méthodes de réduction de dimension, les données ont été **standardisées** par soustraction de la moyenne et division par l'écart-type. Cette étape permet d'assurer une contribution équilibrée de chaque pixel et d'améliorer la stabilité numérique des algorithmes.

5.2.2 Visualisation par PCA

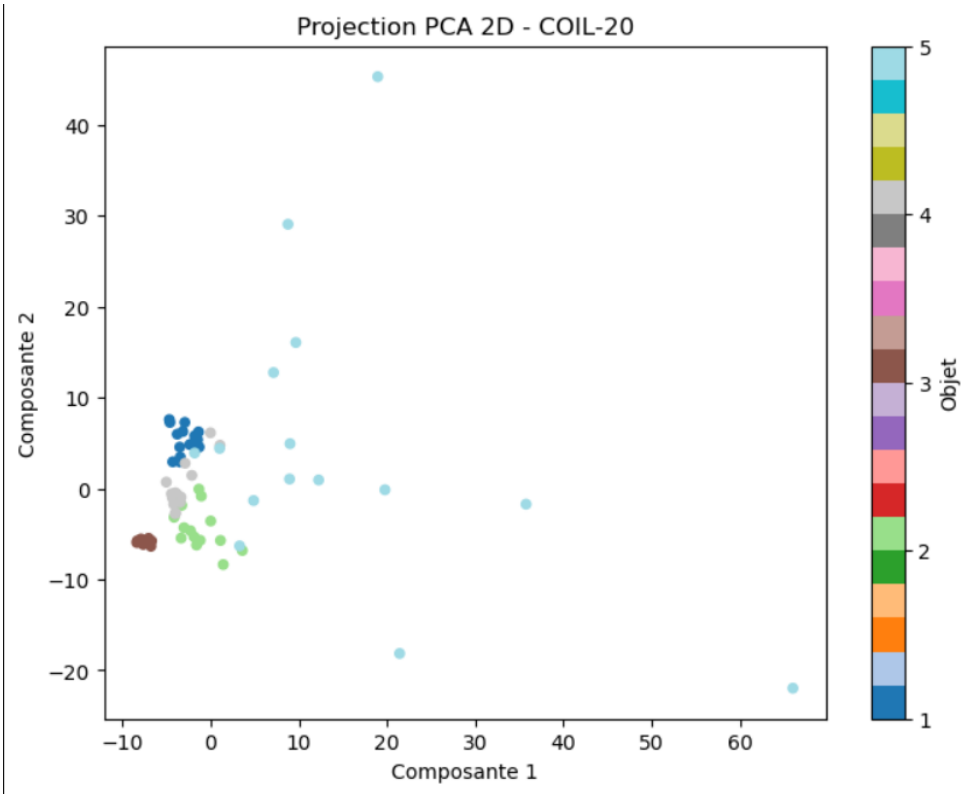

FIGURE 5.6 – Projection PCA en 2D du jeu de données COIL-20

La Figure 5.6 présente la projection bidimensionnelle obtenue par l'analyse en composantes principales (PCA) sur le jeu de données COIL-20.

Interprétation : La PCA projette les données selon les directions de variance maximale, sans exploiter l'information de classe. On observe que certaines classes d'objets forment des regroupements partiels, mais qu'un chevauchement important subsiste entre plusieurs objets.

Ce comportement s'explique par le fait que la variance dominante dans COIL-20 est fortement liée aux variations d'angle de vue, plutôt qu'aux différences de forme entre objets distincts. Ainsi, bien que la PCA permette une réduction de dimension efficace, elle ne garantit pas une séparation optimale des classes dans l'espace projeté.

5.2.3 Visualisation par Isomap

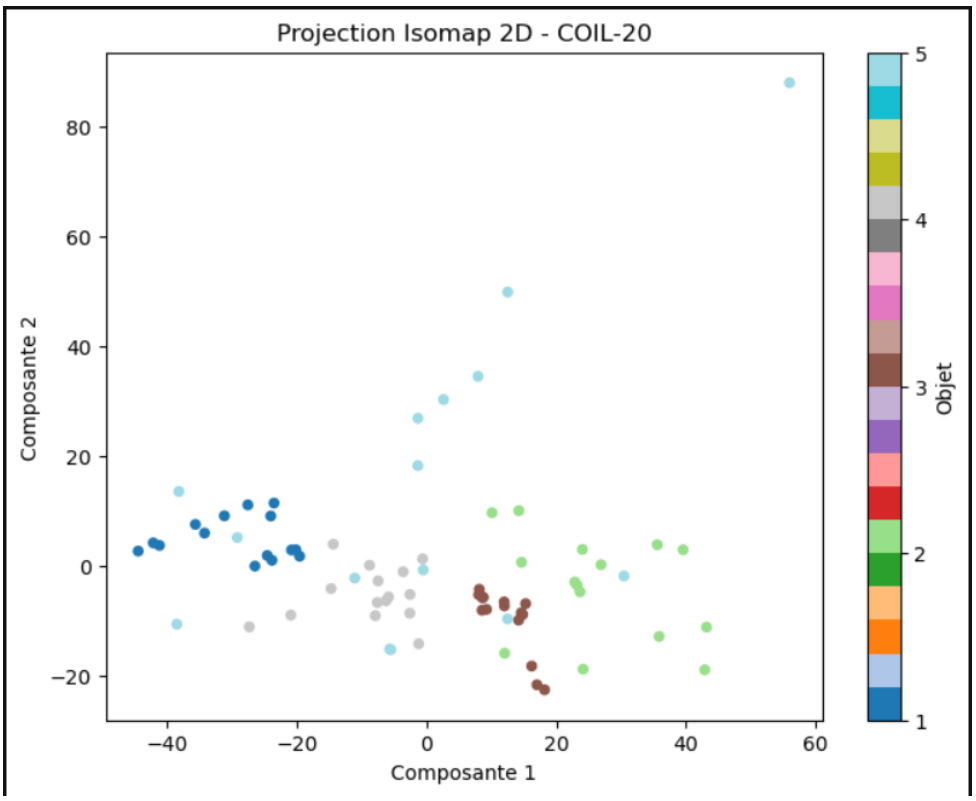

FIGURE 5.7 – Projection Isomap en 2D du jeu de données COIL-20

La Figure 5.7 illustre la projection obtenue à l'aide de la méthode Isomap.

Interprétation : Isomap est une méthode non linéaire qui vise à préserver les distances géodésiques sur la variété sous-jacente des données. La projection met en évidence des structures continues et courbes, correspondant aux variations progressives de l'orientation des objets.

Les images d'un même objet tendent à s'organiser le long de trajectoires lisses, ce qui reflète la nature intrinsèquement non linéaire du jeu de données COIL-20. Toutefois, des recouvrements subsistent entre certains objets présentant des formes visuellement proches.

5.2.4 Visualisation par Laplacian Eigenmaps

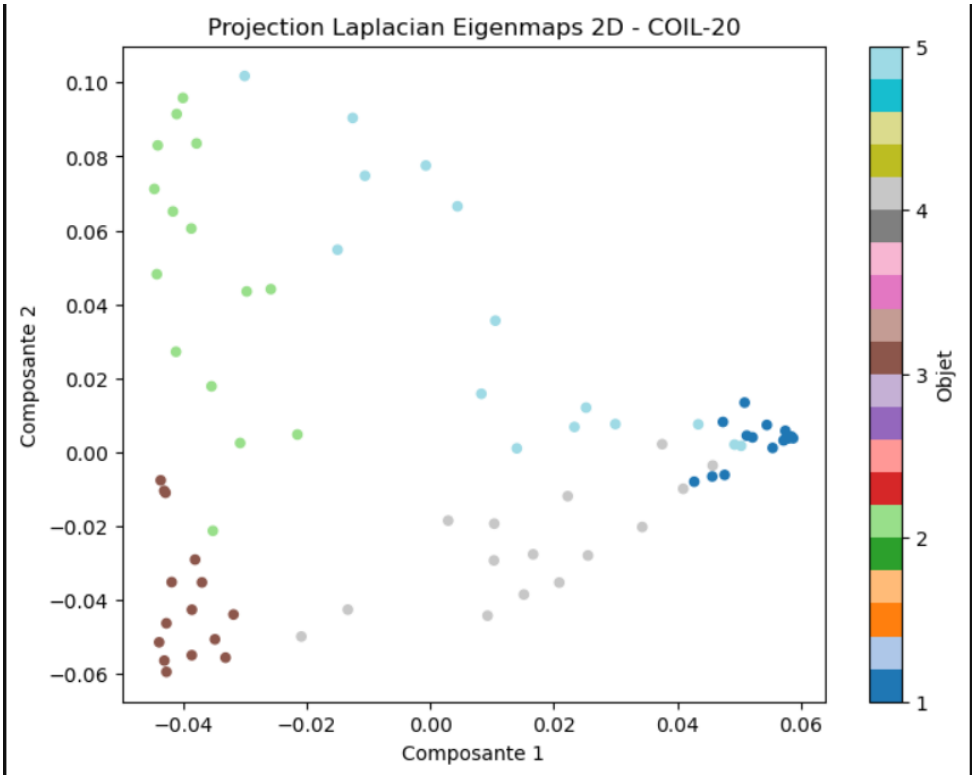

FIGURE 5.8 – Projection Laplacian Eigenmaps en 2D du jeu de données COIL-20

La Figure 5.8 présente la projection obtenue à l'aide de la méthode Laplacian Eigenmaps.

Interprétation : Cette méthode privilégie la préservation des relations locales entre les images voisines dans l'espace original. On observe la formation de regroupements locaux correspondant aux différentes orientations d'un même objet.

Cependant, la séparation globale entre les classes est moins marquée que pour Isomap, ce qui est cohérent avec l'objectif principal de la méthode, axé sur la structure locale plutôt que sur la discrimination inter-classes.

5.2.5 Évaluation des performances de classification

Afin d'évaluer l'impact des projections sur la reconnaissance des objets, un classifieur **KNN** avec $k = 5$ voisins a été appliqué sur les représentations bidimensionnelles obtenues.

Les données ont été séparées en ensembles d'apprentissage et de test selon un ratio de 80% / 20%, en conservant une répartition équilibrée des classes.

Méthode	**Taux de classification**
PCA + KNN	66.67%
Isomap + KNN	73.33%
Laplacian Eigenmaps + KNN	86.67%

TABLE 5.2 – Performances de classification KNN selon la méthode de réduction de dimension pour COIL-20

5.2.6 Discussion des résultats

L'analyse expérimentale menée sur le jeu de données COIL-20 met en évidence l'influence marquée de la structure intrinsèque des données sur l'efficacité des méthodes de réduction de dimension. Contrairement aux données de type chiffres manuscrits, COIL-20 est caractérisé par des variations continues de point de vue pour chaque objet, ce qui induit une organisation naturellement non linéaire des données dans l'espace original.

La PCA, méthode linéaire fondée sur la maximisation de la variance globale, obtient un taux de classification relativement faible de 66.67%. Ce résultat s'explique par le fait que les directions de plus grande variance correspondent principalement aux changements d'orientation des objets, sans garantir une bonne séparation entre les différentes classes. Ainsi, bien que la PCA permette une réduction de dimension efficace, elle ne parvient pas à capturer correctement la structure discriminante du jeu de données COIL-20.

La méthode Isomap améliore les performances de classification, avec un taux de reconnaissance de 73.33%. En cherchant à préserver les distances géodésiques sur la variété sous-jacente, Isomap parvient à mieux modéliser les variations continues liées à la rotation des objets. Cependant, la séparation entre certaines classes reste imparfaite, notamment lorsque des objets présentent des formes ou des silhouettes similaires.

La meilleure performance est obtenue avec la méthode Laplacian Eigenmaps, qui atteint un taux de classification de 86.67%. Cette méthode met l'accent sur la préservation des relations locales entre les images voisines, ce qui s'avère particulièrement pertinent dans le cas de COIL-20, où les images successives d'un même objet sont fortement corrélées. La bonne conservation de la structure locale favorise une représentation plus cohérente des classes, ce qui se traduit par une amélioration significative des performances de classification.

Ces résultats montrent que, pour des données présentant une structure de variété non linéaire dominée par des relations locales, les méthodes basées sur les graphes, et en particulier Laplacian Eigenmaps, constituent une approche plus adaptée que les méthodes linéaires classiques.

5.3 Comparaison des résultats et conclusion générale

Cette section propose une analyse comparative des résultats obtenus lors des deux expériences menées sur les jeux de données MNIST et COIL-20. Elle met en évidence les points forts et les limites de chaque méthode de réduction de dimension, en tenant compte à la fois des performances de classification et de la nature des données traitées.

5.3.1 Comparaison des résultats expérimentaux

Les résultats obtenus montrent que le comportement des méthodes de réduction de dimension dépend fortement de la structure intrinsèque du jeu de données considéré.

Sur le jeu de données MNIST, caractérisé par une variabilité globale importante entre classes, la méthode PCA obtient les meilleures performances de classification. Malgré l'absence d'information de classe, elle parvient à conserver une grande partie de l'information discriminante, ce qui se traduit par un taux de reconnaissance élevé. La LDA, bien que supervisée, est limitée par le faible nombre de dimensions disponibles, ce qui réduit sa capacité de représentation. Les méthodes non linéaires, Isomap et Laplacian Eigenmaps,

offrent des visualisations intéressantes mais présentent des performances de classification inférieures, en partie à cause de leur sensibilité aux paramètres et au bruit.

À l'inverse, sur le jeu de données COIL-20, dont la structure est dominée par des variations continues de point de vue, les méthodes non linéaires se montrent plus adaptées. La PCA, focalisée sur la variance globale, échoue à séparer efficacement les objets, tandis que Isomap améliore la modélisation de la structure géométrique sous-jacente. La meilleure performance est obtenue avec Laplacian Eigenmaps, qui exploite efficacement les relations locales entre images voisines, conduisant à une représentation plus cohérente des classes.

Ces observations confirment qu'une bonne qualité de visualisation ne garantit pas nécessairement de bonnes performances de classification, et que le choix de la méthode doit être guidé par la nature des données et l'objectif visé.

5.3.2 Points forts et limites des méthodes étudiées

La PCA se distingue par sa simplicité, sa rapidité d'exécution et sa robustesse face au bruit. Elle constitue une méthode de référence pour une première exploration des données et pour des tâches de classification à grande échelle. Cependant, son caractère linéaire et non supervisé limite sa capacité à capturer des structures complexes et non linéaires.

La LDA présente l'avantage d'exploiter les informations de classe afin de maximiser la séparabilité entre celles-ci. Elle offre des projections interprétables et efficaces lorsque le nombre de classes est limité. Néanmoins, sa contrainte sur le nombre maximal de dimensions projetées et sa sensibilité aux hypothèses statistiques peuvent restreindre ses performances dans des contextes plus complexes.

Isomap permet de révéler des structures non linéaires globales en préservant les distances géodésiques sur la variété des données. Elle est particulièrement adaptée à des données organisées selon des trajectoires continues, comme dans le cas de COIL-20. Toutefois, son coût computationnel élevé et sa sensibilité au choix des paramètres du graphe constituent des limitations importantes.

Laplacian Eigenmaps met l'accent sur la préservation de la structure locale des données. Cette propriété s'avère très efficace pour des jeux de données présentant de fortes corrélations locales, ce qui explique ses bonnes performances sur COIL-20. En revanche, la méthode ne garantit pas une bonne séparation globale des classes et reste dépendante de la qualité du graphe de voisinage.

Méthode	Points forts	Limites
PCA	— Simple à implémenter — Faible coût computationnel — Robuste au bruit — Efficace pour une première exploration	— Méthode linéaire — Non supervisée — Incapacité à modéliser des structures non linéaires complexes
LDA	— Méthode supervisée — Bonne séparation interclasses — Projections interprétables	— Nombre de dimensions limité à $n\_classes - 1$ — Sensible aux hypothèses statistiques — Moins adaptée aux données complexes
Isomap	— Capture des structures non linéaires globales — Préservation des distances géodésiques — Adaptée aux variétés continues	— Coût computationnel élevé — Sensible au choix du nombre de voisins — Peu robuste au bruit
Laplacian Eigenmaps	— Préservation efficace de la structure locale — Très adaptée aux données fortement corrélées localement — Bon compromis pour la visualisation	— Séparation globale limitée — Dépend fortement de la construction du graphe — Sensible aux paramètres de voisinage

TABLE 5.3 – Comparaison des points forts et des limites des méthodes de réduction de dimension étudiées

Ce tableau synthétise les observations expérimentales et met en évidence le compromis entre simplicité, capacité de représentation et coût computationnel pour chaque méthode.

5.3.3 Conclusion générale

Les expériences réalisées dans ce travail ont permis de mettre en évidence l'importance du choix de la méthode de réduction de dimension en fonction du contexte d'application. Aucune méthode ne peut être considérée comme universellement optimale : chacune présente des avantages spécifiques et des limitations intrinsèques.

Les méthodes linéaires, telles que la PCA et la LDA, se révèlent particulièrement efficaces pour des tâches de classification et pour des jeux de données de grande taille, grâce à leur simplicité et leur robustesse. Les méthodes non linéaires, quant à elles, offrent une meilleure compréhension de la structure géométrique des données et sont mieux adaptées à l'analyse exploratoire et à la visualisation de variétés complexes.

Enfin, ces résultats suggèrent que des approches hybrides, combinant par exemple une

réduction de dimension linéaire suivie d'une méthode non linéaire, constituent une piste intéressante pour concilier performance, interprétabilité et coût computationnel. Ce travail ouvre ainsi la voie à des extensions futures, notamment l'étude de méthodes plus récentes telles que t-SNE ou UMAP, ou l'intégration de techniques de réduction de dimension dans des architectures d'apprentissage profond.
Ce travail montre que la réduction de dimension ne doit pas être considérée comme une étape purement technique, mais comme un choix méthodologique central influençant directement l'interprétation et les performances des modèles.

Bibliographie

[1] I. T. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, 2nd edition, 2002.

[2] R. A. Fisher, *The Use of Multiple Measurements in Taxonomic Problems*, Annals of Eugenics, vol. 7, pp. 179–188, 1936.

[3] J. B. Tenenbaum, V. de Silva, J. C. Langford, *A Global Geometric Framework for Nonlinear Dimensionality Reduction*, Science, vol. 290, no. 5500, pp. 2319–2323, 2000.

[4] M. Belkin, P. Niyogi, *Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering*, Advances in Neural Information Processing Systems (NIPS), 2001.

[5] Y. LeCun, C. Cortes, *MNIST Handwritten Digit Database*, Available at : `http://yann.lecun.com/exdb/mnist/`

[6] S. A. Nene, S. K. Nayar, H. Murase, *Columbia Object Image Library (COIL-20)*, Technical Report CUCS-005-96, Columbia University, 1996, Available at : `https://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php`