Travail de Synthèse Cours Manifold Learning, semestre 1 MALIA, 2024, 1–10
doi: objet numérique sans identifiant (préparation pédagogique)
Date de soumission : octobre 2024
Supervisé par : Dr Jairo Cugliari Duhalde

Étude Comparative Des Algorithmes : PCA, ISOMAP, LLE et Auto-Encoders

Mehdi Mansour[1,*]

[1]Departement Informatique, ICOM, 5 avenue Pierre Mendès France, 69500 Bron, France

[*]Corresponding author. mehdi.mansour@univ-lyon2.fr

Abstract

Cette étude comparative examine les performances de quatre algorithmes de réduction de dimensionnalité: l'Analyse en Composantes Principales (PCA), ISOMAP, l'Apprentissage de Variétés Localement Linéaires (LLE), et les Auto-Encoders. Nous avons évalué ces algorithmes sur des ensembles de données synthétiques (disque, hélice, Swiss roll ondulé) ainsi que sur le jeu de données réel Fashion-MNIST. Notre analyse se concentre sur la capacité de ces méthodes à préserver la structure et l'information des données dans un espace de dimension réduite. Les résultats obtenus corroborent les conclusions de l'état de l'art, démontrant les forces et les limites de chaque approche selon la nature des données traitées. Cette étude, réalisée dans un cadre scolaire, offre un aperçu pratique des différentes techniques de manifold learning et de leur applicabilité dans divers contextes.

Key words: Réduction de dimensionnalité, PCA, ISOMAP, LLE, Auto-Encoders, Manifold Learning, Fashion-MNIST

1 Introduction

La réduction de dimensionnalité est une technique fondamentale en apprentissage automatique et en analyse de données, visant à représenter des données de haute dimension dans un espace de dimension inférieure tout en préservant les caractéristiques essentielles. Cette approche est cruciale pour surmonter le "fléau de la dimensionnalité", améliorer l'efficacité computationnelle, et faciliter la visualisation et l'interprétation des données [1].

Dans cette étude, nous nous concentrons sur quatre méthodes de réduction de dimensionnalité largement utilisées : l'Analyse en Composantes Principales (PCA), ISOMAP, l'Apprentissage de Variétés Localement Linéaires (LLE), et les Auto-Encoders. Chacune de ces méthodes présente des caractéristiques uniques :

- La PCA est une technique linéaire qui projette les données sur les axes de variance maximale [2].
- ISOMAP cherche à préserver les distances géodésiques entre les points dans un espace de dimension réduite [3].
- LLE tente de reconstruire chaque point comme une combinaison linéaire de ses voisins [4].
- Les Auto-Encoders utilisent des réseaux de neurones pour apprendre une représentation compressée des données [5].

Notre objectif est de comparer ces méthodes sur des ensembles de données synthétiques, en évaluant leur capacité à préserver la structure des données et à capturer des relations non linéaires. Cette comparaison vise à fournir des insights sur les forces et les faiblesses de chaque méthode dans différents contextes. Une méthode parmis les quatre sera candidate à effectuer ce travail sur un dataset conséquent du réel, afin de vérifier la robustesse des conclusions lors d'une application sur des données volumineuses et potentiellement bruitées.

2 Matériels et Méthodes

2.1 Ensembles de Données

Notre étude utilise cinq ensembles de données synthétiques soigneusement choisis pour représenter diverses structures de manifolds, ainsi qu'un ensemble de données réelles. Chaque ensemble synthétique a été généré avec et sans bruit gaussien (écart-type $\sigma = 0.05$) pour évaluer la robustesse des méthodes de réduction de dimensionnalité.

2.1.1 Ensembles de Données Synthétiques

2.1.1.1 Hélice Simple

Un manifold 1D enroulé dans un espace 3D, représenté par les équations paramétriques :

$$x = \cos(t), \quad y = \sin(t), \quad z = t/(2\pi) \tag{1}$$

où $t \in [0, 4\pi]$. Cette structure permet d'évaluer la capacité des algorithmes à "dérouler" un manifold non linéaire simple.

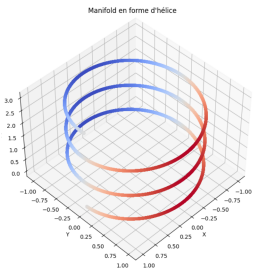

Fig. 0: (a) Hélice

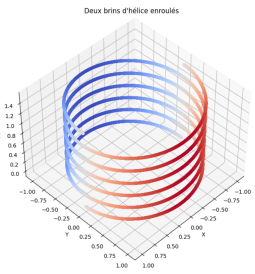

Fig. 0: (b) ADN

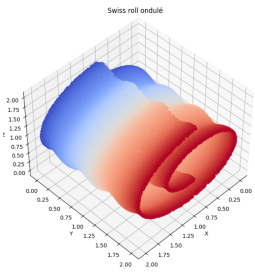

Fig. 0: (c) Swiss Roll

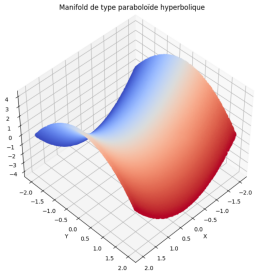

Fig. 0: (d) Selle

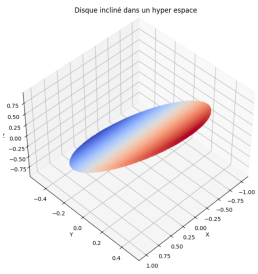

Fig. 0: (e) Disque Incliné

Fig. 1: Visualisation des ensembles de données synthétiques (ensemble des figures ci-dessus

2.1.1.2 Double Hélice

Deux manifolds 1D enroulés, simulant une structure d'ADN, définis par :

$$x_1 = r\cos(t), \quad y_1 = r\sin(t), \quad z_1 = t/(2\pi) \tag{2}$$

$$x_2 = r\cos(t+\pi), \quad y_2 = r\sin(t+\pi), \quad z_2 = t/(2\pi) \tag{3}$$

où $t \in [0, 4\pi]$ et $r = 1$. Ce dataset teste la capacité à distinguer et à préserver deux structures entrelacées.

2.1.1.3 Swiss Roll Ondulé

Un manifold 2D avec une petite ondulation sinusoïdale le long de sa surface, défini par :

$$x = t\cos(t), \quad y = 21 \cdot \text{rand}(0,1), \quad z = t\sin(t) + A\sin(f \cdot y) \tag{4}$$

où $t \in [1.5\pi, 4.5\pi]$, $A = 1$ (amplitude de l'ondulation) et $f = 1$ (fréquence). Cette structure complexe évalue la capacité à "déplier" un manifold non linéaire tout en préservant les relations locales.

2.1.1.4 Paraboloïde Hyperbolique

Un manifold 2D complexe en forme de selle, représenté par :

$$z = \frac{x^2}{a^2} - \frac{y^2}{b^2} \tag{5}$$

où $x, y \in [-2, 2]$ et $a = b = 1$. Cette surface quadrique teste la préservation des courbures positives et négatives dans un espace réduit.

2.1.1.5 Disque Incliné

Un manifold 2D plat mais incliné dans un espace 3D, défini par :

$$x = r\cos(\theta), \quad y = r\sin(\theta)\cos(\phi), \quad z = r\sin(\theta)\sin(\phi) \tag{6}$$

où $r \in [0, 1]$, $\theta \in [0, 2\pi]$, et $\phi = 60°$ (angle d'inclinaison). Ce dataset évalue la capacité à reconnaître et à préserver une structure plane dans un espace 3D.

2.1.2 Ensemble de Données Réelles : Fashion-MNIST

Fashion-MNIST [6] est un ensemble de données de référence composé de 70 000 images en niveaux de gris de 28x28 pixels, représentant 10 catégories de vêtements et d'accessoires. Cet ensemble comprend 60 000 images d'entraînement et 10 000 images de test. Les caractéristiques de Fashion-MNIST sont :

- 10 classes : T-shirt/top, pantalon, pull-over, robe, manteau, sandale, chemise, basket, sac, bottine.
- Chaque image est représentée par un vecteur de 784 dimensions (28x28 pixels).
- Les valeurs des pixels sont normalisées entre 0 et 1.

La Figure 2 présente des exemples d'images de chaque classe de Fashion-MNIST.

Fig. 2: Échantillons d'images de l'ensemble de données Fashion-MNIST, illustrant les 10 classes différentes.

L'utilisation de Fashion-MNIST permet d'évaluer les performances des méthodes de réduction de dimensionnalité sur des données réelles de haute dimension, tout en offrant un contexte d'application concret dans le domaine de la vision par ordinateur et de la classification d'images.

La diversité de ces ensembles de données, allant de structures géométriques simples à des manifolds complexes et des données

réelles, offre un cadre complet pour évaluer les forces et les faiblesses de chaque méthode de réduction de dimensionnalité étudiée.

2.2 Algorithmes utilisés

Dans cette étude, nous avons implémenté et comparé quatre méthodes de réduction de dimensionnalité, chacune ayant ses propres fondements mathématiques et promesses théoriques :

2.2.1 Analyse en Composantes Principales (PCA)

2.2.1.1 Fondements mathématiques :

La PCA est une technique linéaire qui projette les données sur les axes de variance maximale. Mathématiquement, pour un ensemble de données X de dimension $n \times p$, la PCA cherche à trouver une matrice de projection W telle que :

$$Y = XW \tag{7}$$

où Y est la représentation réduite des données. W est obtenue en résolvant le problème aux valeurs propres :

$$C_X W = W\Lambda \tag{8}$$

où C_X est la matrice de covariance de X et Λ est une matrice diagonale des valeurs propres.

2.2.1.2 Promesses théoriques :

- Préservation optimale de la variance globale des données.
- Décorrélation des caractéristiques dans l'espace réduit.
- Garantie de reconstruction optimale au sens des moindres carrés.

2.2.2 ISOMAP

2.2.2.1 Fondements mathématiques :

ISOMAP étend l'idée du Multi-Dimensional Scaling (MDS) en remplaçant les distances euclidiennes par des distances géodésiques estimées. Le processus se déroule en trois étapes :

1) Construction d'un graphe des k plus proches voisins. 2) Calcul des distances géodésiques entre tous les points en utilisant l'algorithme de Dijkstra. 3) Application du MDS classique sur la matrice des distances géodésiques.

La fonction objectif à minimiser est :

$$E = \sum_{i<j} (d_G(x_i, x_j) - \|y_i - y_j\|)^2 \tag{9}$$

où d_G est la distance géodésique et y_i, y_j sont les coordonnées dans l'espace réduit.

2.2.2.2 Promesses théoriques :

- Préservation de la structure globale non linéaire des données.
- Convergence asymptotique vers la vraie structure du manifold sous certaines conditions.

2.2.3 Locally Linear Embedding (LLE)

2.2.3.1 Fondements mathématiques :

LLE reconstruit chaque point comme une combinaison linéaire de ses voisins. Le processus comprend deux étapes principales :

1) Calcul des poids de reconstruction W_{ij} en minimisant :

$$\epsilon(W) = \sum_i \|x_i - \sum_j W_{ij} x_j\|^2 \tag{10}$$

2) Calcul des coordonnées de basse dimension Y en minimisant :

$$\Phi(Y) = \sum_i \|y_i - \sum_j W_{ij} y_j\|^2 \tag{11}$$

sous contraintes de centrage et d'échelle.

2.2.3.2 Promesses théoriques :

- Préservation des relations locales entre les points.
- Invariance aux rotations, translations et mises à l'échelle.

2.2.4 Auto-Encodeurs

2.2.4.1 Fondements mathématiques :

Les auto-encodeurs sont des réseaux de neurones qui apprennent une représentation compressée (encodage) des données d'entrée, puis tentent de reconstruire ces données (décodage). Formellement, pour une entrée x, l'auto-encodeur apprend une fonction d'encodage f et une fonction de décodage g telles que :

$$\hat{x} = g(f(x)) \tag{12}$$

où $\hat{x}$ est la reconstruction de x. L'objectif est de minimiser l'erreur de reconstruction :

$$L(x, \hat{x}) = \|x - \hat{x}\|^2 \tag{13}$$

2.2.4.2 Promesses théoriques :

- Capacité à capturer des relations hautement non linéaires.
- Flexibilité dans la conception de l'architecture pour s'adapter à différents types de données.
- Possibilité d'intégrer des contraintes spécifiques au domaine dans le processus d'apprentissage.

Chacune de ces méthodes offre une approche unique pour aborder le problème de la réduction de dimensionnalité, avec des forces et des limitations spécifiques que nous explorerons à travers nos expériences sur les datasets synthétiques et réels.

3 Protocole Expérimental

Notre étude comparative a suivi un protocole expérimental conçu pour évaluer équitablement les performances de chaque méthode de réduction de dimensionnalité sur différents ensembles de données, Voici les détails de notre approche :

3.1 Préparation des Données

3.1.1 Ensembles de Données Synthétiques

Pour la génération des datasets synthétiques, nous avons procédé comme suit :

- Génération des coordonnées 3D initiales selon les équations spécifiques à chaque manifold.
- Enrichissement de 97 dimensions supplémentaires avec des coordonnées tirées aléatoirement d'une distribution normale, puis normalisées pour garder une variance acceptable qui ne phagocyte pas la variance du vrai manifold.
- Création de deux versions de chaque dataset : une sans bruit et une avec un bruit gaussien (écart-type $\sigma = 0.05$) ajouté aux 100 dimensions.

Cette approche nous a permis de simuler des données de haute dimension tout en préservant la structure sous-jacente du manifold original.

3.1.2 Ensemble de Données Réelles : Fashion-MNIST

Pour notre étude sur des données réelles, nous avons utilisé le dataset Fashion-MNIST [6] avec des auto encodeurs convolutifs. Ce choix offre un bon compromis entre complexité et taille de données:

- **Prétraitement** :
  - Normalisation : Les valeurs des pixels ont été normalisées entre 0 et 1 en divisant par 255.
  - Reshape : Les images ont été redimensionnées de (28, 28) à (28, 28, 1) pour être compatibles avec l'architecture d'entrée des auto encodeurs convolutifs.

3.2 Infrastructure de Calcul

Les expérimentations ont souvent dû être répétées et certains calculs étaient conséquents. Le risque était de bloquer l'avancement du travail ou de limiter le champ des explorations à des valeurs très réduites et optimisées de manière caricaturale. Pour contourner cet inconvénient, le choix a été de s'appuyer autant que possible sur des infrastructures matérielles et des algorithmes permettant d'accélérer les calculs :

- **GPU NVIDIA A100** : Accélération matérielle pour les opérations de calcul matriciel et l'entraînement des réseaux de neurones.
- **Librairies du projet "RAPIDS.ai"** : Implémentation GPU-accélérée des algorithmes standards de machine learning et de calcul matriciel, telles que cuML comme substitut rapide de scikit-learn ou cupy en alternative à numpy.
- **TensorFlow avec support GPU** : Utilisé pour l'implémentation et l'entraînement des auto encodeurs.
- **Alternatives sur CPUs multicœurs** : Certains algorithmes n'avaient pas d'implémentation compatible avec l'accélération sur GPU, comme ISOMAP. Nous avons utilisé un maximum de cœurs de CPU en définissant le paramètre `n_jobs=-1` dans les méthodes qui le permettaient.

3.3 Optimisation des Hyperparamètres

Dans le cadre de notre étude comparative des méthodes de réduction de dimensionnalité, le calibrage des hyperparamètres est une étape cruciale pour garantir des performances optimales de chaque algorithme. Traditionnellement, des techniques telles que la validation croisée avec recherche sur grille (*Grid Search*) ou aléatoire (*Random Search*) sont utilisées. Cependant, ces approches peuvent être inefficaces en termes de temps de calcul et ne sont pas adaptées aux espaces de recherche de grande dimension ou complexes.

Pour surmonter ces limitations, nous avons opté pour l'utilisation d'Optuna [7], un framework d'optimisation d'hyperparamètres automatisé et efficace. Optuna permet une exploration intelligente de l'espace des hyperparamètres en utilisant des techniques d'optimisation bayésienne, notamment l'algorithme TPE (*Tree-structured Parzen Estimator*).

3.3.1 Pourquoi Optuna ?

Optuna offre plusieurs avantages qui le rendent particulièrement adapté à notre étude :

- **Optimisation efficace** : Utilise des méthodes d'optimisation séquentielle basée sur des modèles (SMBO), ce qui permet d'explorer l'espace des hyperparamètres de manière plus efficace que les recherches exhaustives.
- **Flexibilité** : Permet de définir des espaces de recherche complexes, incluant des hyperparamètres discrets, continus et catégoriels.
- **Pruning adaptatif** : Peut interrompre précocement les essais non prometteurs, économisant ainsi des ressources computationnelles.
- **Intégration facile** : S'intègre aisément avec les frameworks de machine learning courants tels que Scikit-learn et TensorFlow.

3.3.2 Fondements Mathématiques d'Optuna

L'algorithme principal utilisé par Optuna pour l'optimisation des hyperparamètres est le TPE (*Tree-structured Parzen Estimator*). Le TPE est une méthode d'optimisation bayésienne qui modélise la fonction objectif en utilisant des estimations de densité de probabilité.

3.3.2.1 Principe du TPE

Le TPE cherche à modéliser la probabilité inverse $p(x|y)$ plutôt que la probabilité directe $p(y|x)$. Pour ce faire, il divise les observations en deux groupes :

- $l(x)$: Distribution des hyperparamètres ayant conduit à des performances "bonnes", c'est-à-dire avec une valeur de la fonction objectif inférieure à un seuil y^*.
- $g(x)$: Distribution des hyperparamètres ayant conduit à des performances "mauvaises", c'est-à-dire avec une valeur de la fonction objectif supérieure ou égale à y^*.

L'objectif est de choisir les hyperparamètres x qui maximisent le rapport d'acquisition :

$$\mathrm{EI}(x) = \frac{l(x)}{g(x)} \tag{14}$$

où $\mathrm{EI}(x)$ est l'*Expected Improvement* (amélioration espérée).

3.3.2.2 Processus d'Optimisation

Le processus se déroule en plusieurs étapes :

1. **Définition de l'espace de recherche** : L'utilisateur spécifie les hyperparamètres à optimiser et leurs domaines respectifs.
2. **Échantillonnage** : Optuna propose un ensemble d'hyperparamètres en maximisant le rapport $\frac{l(x)}{g(x)}$.
3. **Évaluation** : Le modèle est entraîné avec ces hyperparamètres et la performance est évaluée.
4. **Mise à jour des distributions** : Les distributions $l(x)$ et $g(x)$ sont mises à jour avec les nouvelles observations.
5. **Itération** : Le processus est répété jusqu'à atteindre un critère d'arrêt (nombre maximal d'essais, convergence, etc.).

3.3.2.3 Pruning Adaptatif

Optuna intègre également des techniques de *pruning* adaptatif, basées sur des algorithmes tels que Successive Halving et Hyperband. Le principe est d'arrêter précocement les essais dont les performances intermédiaires sont inférieures à celles attendues,

permettant ainsi de concentrer les ressources sur les configurations prometteuses.

3.3.3 Application à nos Méthodes

Pour chaque algorithme de réduction de dimensionnalité, nous avons défini un espace de recherche spécifique :

- **PCA** : Optimisation du nombre de composantes principales $n_{\text{components}} \in [2, 10]$.
- **ISOMAP** : Optimisation du nombre de voisins $n_{\text{neighbors}} \in [4, 30]$ et du nombre de composantes $n_{\text{components}} \in [2, 10]$.
- **LLE** : Optimisation du nombre de voisins $n_{\text{neighbors}} \in [5, 30]$ et du nombre de composantes $n_{\text{components}} \in [2, 10]$.
- **Auto-Encodeurs** : Optimisation du nombre de couches $n_{\text{layers}} \in [2, 5]$, du nombre de neurones par couche $n_{\text{neurons}} \in \{64, 128, 256\}$, du taux de drop-out dropout_rate $\in [0.0, 0.5]$ et de la taille de la couche latente $n_{\text{components}} \in [2, 10]$.

Aussi, La fonction objectif à minimiser ou maximiser a été adaptée à chaque méthode :

- **PCA** : Maximisation de la variance expliquée cumulée par les composantes retenues.
- **ISOMAP et LLE** : Minimisation de l'erreur de reconstruction géodésique ou locale.
- **Auto-Encodeurs** : Minimisation de l'erreur de reconstruction (MSE) entre les données originales et reconstruites.

3.3.4 Critères d'Arrêt et Paramètres d'Optimisation

Nous avons fixé un nombre maximal d'essais (50 essais par étude) pour l'optimisation, en veillant à trouver un équilibre entre le temps de calcul et la qualité de l'optimisation. Le *pruning* adaptatif d'Optuna a permis de réduire le temps de calcul en arrêtant précocement les configurations non prometteuses.

3.4 Procédure d'Évaluation

Pour évaluer les performances de chaque méthode de réduction de dimensionnalité avec les hyperparamètres optimisés, nous avons suivi une procédure en plusieurs étapes :

3.4.1 Récupération des coordonnées dans l'espace réduit

Chaque méthode a été appliquée sur les ensembles de données synthétiques et réels en utilisant les meilleurs hyperparamètres trouvés par Optuna. Cela garantit que chaque algorithme est évalué dans ses conditions optimales.

3.4.2 Évaluation Quantitative

Nous avons mesuré plusieurs métriques pour quantifier les performances :

- **Erreurs de Reconstruction** : Dans le cas d'Isomap par exemple, l'erreur de reconstruction est calculée en comparant la matrice des distances géodésiques centrées dans l'espace original et les valeurs propres obtenues dans l'espace réduit. La formule utilisée est :

$$E = \frac{\sqrt{\sum(G^2_{\text{center}}) - \sum(\lambda^2)}}{n} \tag{15}$$

où G_{center} est la matrice des distances géodésiques centrée dans l'espace d'origine, et λ sont les valeurs propres résultantes de la décomposition KernelPCA dans l'espace réduit. Cette erreur quantifie dans quelle mesure l'embedding réduit capture la structure géométrique des distances dans l'espace d'origine.

- **Préservation des Distances Globales** : Évaluée à l'aide de la mesure de stress géodésique, qui quantifie la distorsion des distances globales lors de la réduction de dimensionnalité. Cette mesure est particulièrement adaptée pour des méthodes comme **Isomap**, qui cherchent à préserver les distances géodésiques sur le manifold. Le stress est calculé comme suit :

$$S_{\text{Isomap}} = \sqrt{\frac{\sum_{i<j}(d_{ij}^{\text{geo}} - d_{ij}^{\text{low}})^2}{\sum_{i<j}(d_{ij}^{\text{geo}})^2}} \tag{16}$$

où d_{ij}^{geo} est la distance géodésique dans l'espace de haute dimension, et d_{ij}^{low} est la distance dans l'espace réduit. Un stress faible indique une bonne préservation des distances géodésiques globales et de la structure globale du manifold.

- **Préservation des Distances Locales** : Évaluée à l'aide d'une mesure de stress local, qui quantifie la distorsion des distances entre un point et ses voisins immédiats. Cette mesure est adaptée pour des méthodes comme **LLE**, qui visent à préserver les relations locales entre les points. Le stress local est calculé comme suit :

$$S_{\text{LLE}} = \sqrt{\frac{\sum_{i<j, j\in N(i)}(d_{ij}^{\text{high}} - d_{ij}^{\text{low}})^2}{\sum_{i<j, j\in N(i)}(d_{ij}^{\text{high}})^2}} \tag{17}$$

où d_{ij}^{high} est la distance dans l'espace de haute dimension, et d_{ij}^{low} est la distance dans l'espace réduit pour les points j dans le voisinage immédiat $N(i)$ de i. Un stress faible indique une bonne préservation des relations locales entre les points voisins.

- **Performance de Classification** : Pour l'ensemble de données Fashion-MNIST, nous avons entraîné un classifieur de type regréssion logistique multinomiale sur les données en espace réduit et non-réduit et comparé l'exactitude des prédiction.

3.4.3 Évaluation Qualitative

Nous avons également réalisé une analyse visuelle :

- **Visualisation des Projections** : Création de graphiques en 2D et 3D pour visualiser la répartition des données dans l'espace réduit, ce qui permet d'évaluer la séparation des classes et la préservation des structures globales.
- **Analyse des Manifolds** : Observation de la capacité des méthodes à "déplier" les structures non linéaires, telles que le Swiss Roll ou l'hélice, et à préserver les voisinages locaux.

3.4.4 Comparaison Globale

Enfin, nous avons comparé les méthodes entre elles en tenant compte des différentes métriques, afin d'identifier les forces et les faiblesses de chaque approche dans différents contextes.

4 Résultats

4.1 Performances sur les Ensembles de Données Synthétiques

4.1.1 Analyse en Composantes Principales (PCA)

La projection des observations dans l'espace réduit est illustrée dans la Figure 8 de l'appendice.

La PCA a montré des performances cohérentes sur tous les ensembles de données, avec une variance expliquée relativement élevée pour les manifolds linéaires comme le disque incliné. Cependant, elle a eu des difficultés à capturer la structure non linéaire du Swiss roll ondulé et .

Table 1. Somme des variances expliquées (%) pour les 2 premiers axes PCA

Dataset	Type	Sans bruit	Avec bruit
Hélice	1d	75.87	75.80
Double hélice	1d*2	84.15	83.93
Swiss roll ondulé	2	73.64	72.36
Paraboloïde hyperbolique	2	75.83	76.63
Disque incliné	2d	99.97	99.26

Nous remarquons la redoutable efficacité de la PCA sur des données linéaires comme le disque dont elle arrive à capturer la quasi totalité de la variance dans les premières composantes.

Cette méthode s'est également montrée particulièrement robuste face au bruit, les datasets bruitées sont réduits avec presque autant d'efficacité que les datasets sans bruit.

4.1.2 ISOMAP (Isometric Feature Mapping)

La projection des observations dans l'espace réduit par ISOMAP est illustrée dans la Figure 9 de l'appendice.

ISOMAP a démontré une capacité supérieure à détecter la structure globale de manifolds non linéaires,

Table 2. Erreurs de reconstruction ISOMAP (basées sur les distances géodésiques globales)

Dataset	Type	Sans bruit	Avec bruit
Hélice	1d	0.0002	0.0158
Double hélice	1d*2	0.4085	0.3455
Swiss roll ondulé	2d	0.0975	0.1394
Paraboloïde hyperbolique	2d	0.3898	0.4114
Disque incliné	2d	0.0027	0.0060

ISOMAP démontre une efficacité remarquable à capturer ces structures non linéaires et à les visualiser, en particulier la parabloîde hyperbolique et le suiss roll ondulé. Néanmois la technique se montre sensible au bruit qui baisse significativement sa performance sur le swiss roll par exemple.

Le Swiss roll ondulé, structure non linéaire complexe, est relativement bien capturé par ISOMAP avec une erreur de 0.0975 sans bruit, augmentant légèrement à 0.1394 avec bruit. Cette performance est nettement meilleure que celle observée avec la PCA pour cette structure.

ISOMAP montre des limitations pour la structure en double hélice qu'elle traite similairement à la PCA.

Le disque incliné, bien que linéaire, est très bien capturé par ISOMAP, avec une erreur minime de 0.0027 sans bruit, augmentant légèrement à 0.0060 avec bruit.

La méthode montre également une robustesse variable au bruit, dépendant fortement de la nature intrinsèque des données analysées.

4.1.3 Locally Linear Embedding LLE

La projection des observations dans l'espace réduit par LLE est illustrée dans la Figure 10 de l'appendice.

LLE a excellé dans la préservation des relations locales, notamment en trouvant la bonne dimension intrinsèque de l'hélice et de la double hélice. Elles sont en effet ramenés à une unique dimension, avec les deux brins de la double hélice qui sont bien différenciés.

Table 3. Erreurs de reconstruction LLE (basées sur distances locales entre paires de points)

Dataset	Type	Sans bruit	Avec bruit
Hélice	1d	1.6098	1.6122
Double hélice	1d*2	1.2987	1.3043
Swiss roll ondulé	2d	1.1692	1.1771
Paraboloïde hyperbolique	2d	2.9057	2.8753
Disque incliné	2d	0.6131	0.6081

LLE se montre robuste au bruit sauf pour la structure en double hélice dont le bruit cause une confusion dans la séparation des brins.

Comparé à ISOMAP et PCA, LLE montre clairement une capacité à mieux reconnaitre un motif enroulé et à potentiellement mieux séparer des manifolds entrelacés.

4.1.4 Auto-Encodeurs

La projection des observations dans l'espace réduit par les Auto-Encoders est illustrée dans la Figure 11 de l'appendice.

Les Auto-Encodeurs semblent concentrer les caractéristiques essentielles de chaque manifold dans un espace de moindre dimension. Sur l'hélice et la double hélice nous remarquons la sauvegarde du nombre de tours et du sens de rotation des brins.

Table 4. Erreurs de reconstruction des Auto-Encodeurs, MSE (entre originaux et reconstruits à partir de l'encodage)

Dataset	Type	Sans bruit	Avec bruit
Hélice	1d	0.0140	0.0141
Double hélice	1d*2	0.0055	0.0054
Swiss roll ondulé	2d	0.0046	0.0049
Paraboloïde hyperbolique	2d	0.0347	0.0332
Disque incliné	2d	0.0019	0.0018

Les Auto-Encodeurs montrent ainsi des performances exceptionnelles dans la capture des caractéristiques des données :

- **Efficacité globale :** Les erreurs de reconstruction sont remarquablement faibles pour tous les datasets, allant de 0.0019 à 0.0347 sans bruit, indiquant une excellente capacité à capturer l'essence des structures de données variées.
- **Structures non linéaires complexes :** Le Swiss roll ondulé (erreur de 0.0046 sans bruit) et la double hélice (0.0055 sans bruit) sont particulièrement bien reconstruits, démontrant l'aptitude des Auto-Encodeurs à saisir des relations non linéaires complexes.
- **Adaptabilité :** Bien que l'erreur soit légèrement plus élevée pour le paraboloïde hyperbolique (0.0347 sans bruit), elle reste très faible, illustrant la capacité d'adaptation des Auto-Encodeurs à diverses géométries.

La robustesse au bruit des Auto-Encodeurs est exceptionnelle:

- Les erreurs de reconstruction restent pratiquement inchangées en présence de bruit pour tous les datasets.
- Dans certains cas, comme pour le disque incliné et le paraboloïde hyperbolique, l'erreur diminue même légèrement avec le bruit (de 0.0019 à 0.0018 et de 0.0347 à 0.0332 respectivement), suggérant une possible régularisation induite par le bruit.

Ces résultats mettent en évidence plusieurs points forts des Auto-Encodeurs :

1. **Flexibilité :** Ils s'adaptent efficacement à une grande variété de structures de données, des plus simples aux plus complexes.
2. **Robustesse :** Leur performance reste stable même en présence de bruit, indiquant une grande robustesse.
3. **Capture de caractéristiques :** Les faibles erreurs de reconstruction sur tous les datasets suggèrent une excellente capacité à capturer les caractéristiques essentielles des données, même dans des espaces de dimension réduite.

Comparés aux autres méthodes (PCA, ISOMAP, LLE), les Auto-Encodeurs semblent offrir la meilleure performance globale et la plus grande polyvalence. Ils combinent la capacité de PCA à traiter efficacement les structures linéaires, la flexibilité d'ISOMAP pour les manifolds non linéaires, et la robustesse au bruit de LLE, tout en surpassant ces méthodes en termes de précision de reconstruction.

Cette polyvalence et cette efficacité font des Auto-Encoders un outil puissant pour la réduction de dimensionnalité, particulièrement adapté aux ensembles de données complexes et potentiellement bruités. Leur capacité à capturer des caractéristiques subtiles des données en fait également un choix intéressant pour des tâches d'extraction de caractéristiques et de représentation d'apprentissage.

Cette complexité a bien sûr un coût calculatoire car la complexité de l'architecture des auto-encodeurs induit une stratification d'opérations matricielles ardues, mais heureusement parallélisable sur certaines configurations.

Notons que nous avons écarté la comparaison des temps de calcul des différentes méthodes, faute d'avoir pu leur offrir les mêmes conditions d'exécution. En effet, Par manque de temps, chaque méthode a été appliquée avec la configuration la plus rapide possible.

4.1.5 Application des Auto-Encodeurs aux données Fashion MNIST

Après avoir étudié les performances des Auto-Encodeurs sur des datasets synthétiques, nous avons appliqué cette méthode à un ensemble de données réelles : Fashion MNIST. Cette application a mis au défi la capacité des Auto-Encodeurs à traiter des données complexes et de haute dimension dans un contexte réel.

4.1.5.1 Configuration et Optimisation

Nous avons utilisé un Auto-Encodeur convolutif, optimisé avec Optuna pour trouver les meilleurs hyperparamètres. Cette approche a permis d'adapter l'architecture du réseau à la complexité spécifique des images de vêtements.

4.1.5.2 Réduction de Dimensionnalité

L'Auto-Encodeur a réduit efficacement la dimensionnalité des images de 784 (28x28 pixels) à un espace latent de dimension 7 fois inférieure, tout en préservant les caractéristiques essentielles pour la reconstruction et la classification.

4.1.5.3 Performances

- **Reconstruction sur le set de test** : L'erreur moyenne de reconstruction (MSE) était de 0.0031, démontrant la capacité du modèle à capturer et à reconstruire fidèlement les caractéristiques des images.

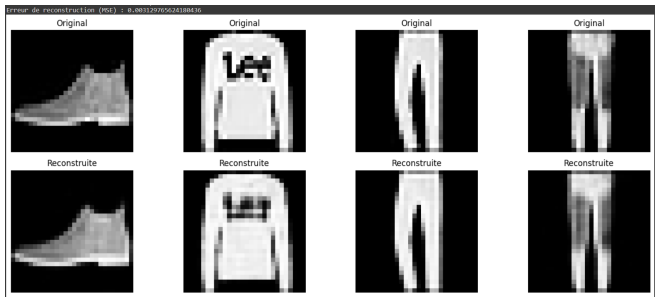

Fig. 3: Exemples de reconstruction des images sur le set de test par l'autoencodeur

- **Conservation de l'Information** : La visualisation t-SNE en 2d des représentations latentes et des représentations d'origine a révélé une séparation meilleure des différentes catégories de vêtements dans la représentation réduite, indiquant que l'information discriminante était bien préservée et bien captée dans l'espace réduit, en particulier pour la classe "pantalon" qui semble avoir des caractéristiques uniques.

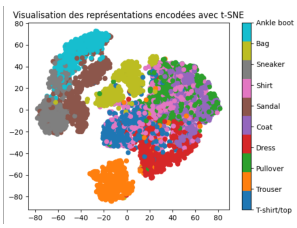

Fig. 4: t-SNE des représentations latentes

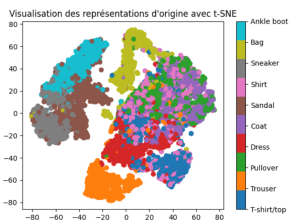

Fig. 5: t-SNE des données originales

- **Comparaison avec PCA** : Les projections PCA des données originales et encodées confirment que l'Auto-Encodeur a capturé dans un espace de l'ordre de 10 fois moindre, l'essentiel de l'information de séparabilité des classes. la distribution des classes dans les 3 premières composantes principales est plus claire dans l'espace réduit.

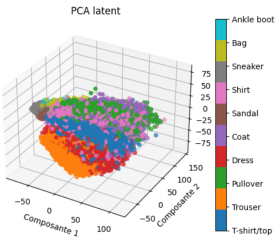

Fig. 6: PCA des représentations latentes

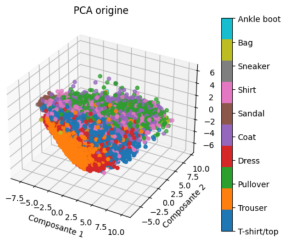

Fig. 7: PCA des données originales

4.1.5.4 Évaluation de la Qualité de la Réduction

Pour évaluer la pertinence de l'information conservée, nous avons comparé les performances d'un classifieur (régression logistique) sur les données originales et sur les représentations latentes. Les résultats pour un nombre maximal d'itérations de 500 sont résumés dans le tableau ci-dessous :

Métrique	Origine	Latent
Accuracy	84.38%	85.00%
Precision	84.25%	84.84%
Recall	84.38%	85.00%
F1-Score	84.29%	84.89%

Table 5. Comparaison des performances de la régression logistique sur l'espace d'origine et l'espace latent, $\max_i ter = 500$

Ces résultats montrent que l'Auto-Encodeur a réussi à préserver, voire améliorer légèrement, l'information essentielle pour la classification, malgré une réduction significative de la dimensionnalité. La précision sur l'espace latent (85.00%) dépasse légèrement celle sur l'espace d'origine (84.38%), suggérant que la compression réalisée par l'Auto-Encodeur permet une représentation plus efficace de l'information pertinente pour la tâche de classification. Il est raisonnable d'affirmer que l'Auto-Encodeur a permis une compression sans perte significative de qualité pour la classification, et qu'il a amélioré légèrement la qualité de la prédiction, probablement grâce à l'élimination du bruit à la classification , c'est à dire les caractéristiques non descriminantes des classes.

5 Discussion

Notre étude comparative a mis en lumière les forces et les faiblesses de chaque méthode de réduction de dimensionnalité :

- **PCA** s'est révélée efficace pour les structures linéaires et robuste au bruit, mais limitée pour capturer des relations non linéaires complexes.
- **ISOMAP** a excellé dans la préservation de la structure globale des manifolds non linéaires, mais a montré des limitations pour les données très bruitées ou les structures très complexes.
- **LLE** a démontré une forte capacité à préserver les relations locales, mais s'est avérée sensible au bruit et moins performante pour les structures globales complexes.
- Les **Auto-Encoders** ont montré une grande adaptabilité et des performances solides sur une variété de structures de données, mais au prix d'un temps de calcul plus élevé et d'une complexité accrue dans le réglage des hyperparamètres.

Ces résultats corroborent largement les conclusions de la littérature existante [1, 8]. Ils soulignent l'importance de choisir la méthode de réduction de dimensionnalité en fonction de la nature spécifique des données et des objectifs de l'analyse.

L'expérience sur Fashion-MNIST a particulièrement mis en évidence la capacité des Auto-Encoders à préserver l'information discriminante dans un contexte de classification d'images, suggérant leur pertinence pour les tâches de vision par ordinateur.

6 Conclusion

Cette étude comparative offre un aperçu pratique des performances de quatre méthodes de réduction de dimensionnalité largement utilisées. Nos résultats soulignent qu'il n'existe pas de méthode universellement supérieure, mais que le choix dépend fortement de la structure des données et des objectifs spécifiques de l'analyse.

Les perspectives futures incluent l'exploration de méthodes plus récentes comme t-SNE ou UMAP, ainsi que l'investigation de l'impact de la réduction de dimensionnalité sur d'autres tâches d'apprentissage automatique au-delà de la classification.

Cette étude, dans un cadre de consolidation des connaissances en fin du module Manifold learning, a clairement contribué à une meilleure compréhension des forces et des limites des différentes approches de réduction de la dimensionnalité expérimentées.

References

1. Laurens Van der Maaten, Eric Postma, and Jaap Van den Herik. Dimensionality reduction: a comparative review. *J Mach Learn Res*, 10(66-71):13, 2009.
2. Ian T Jolliffe. *Principal component analysis*. Springer, 2002.
3. Joshua B Tenenbaum, Vin De Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.
4. Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.
5. Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *science*, 313(5786):504–507, 2006.
6. Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.
7. Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.
8. Fuzhen Wang, Minghui Jiang, Chen Qian, Shuo Yang, Cheng Li, Honggang Zhang, Xiaogang Wang, and Xiaoou Tang. A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1):43–76, 2021.

Appendices

A Visualisations 2D des espaces réduits des données simulées

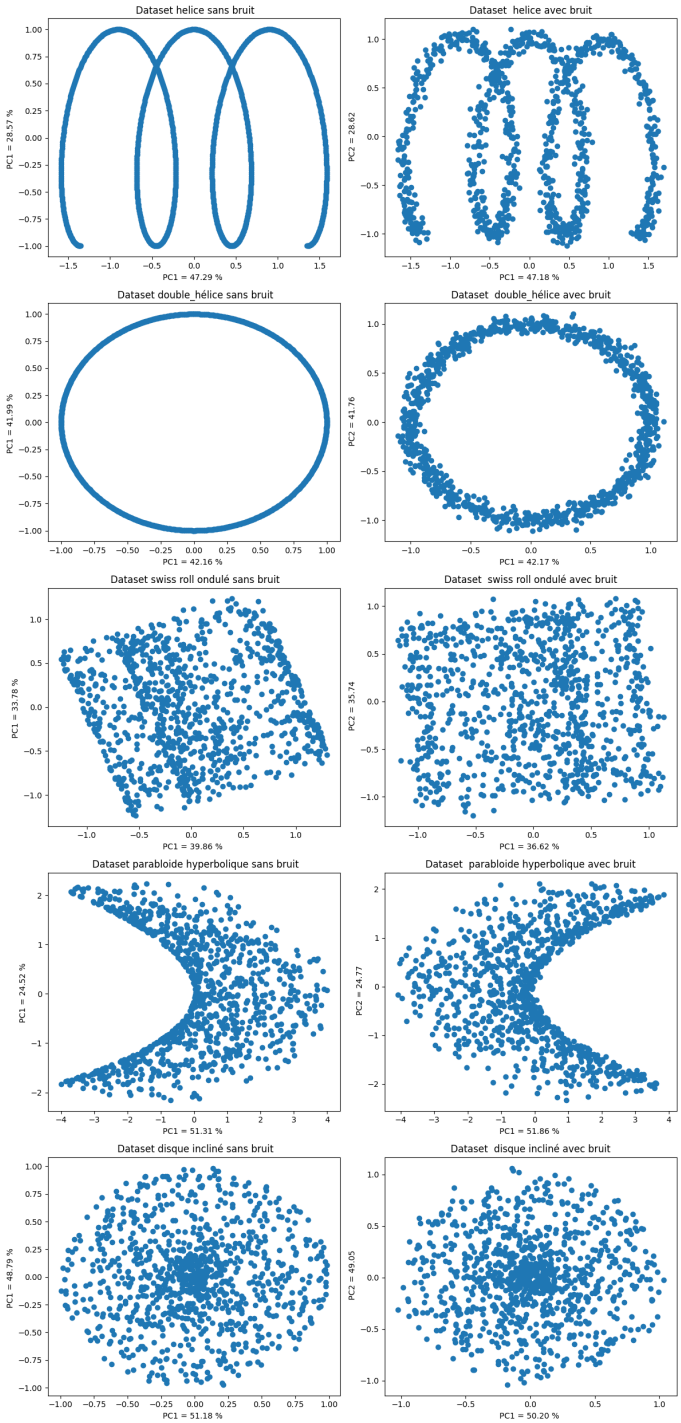

Fig. 8: Résultats de la méthode PCA sur les ensembles de données synthétiques avec et sans bruit.

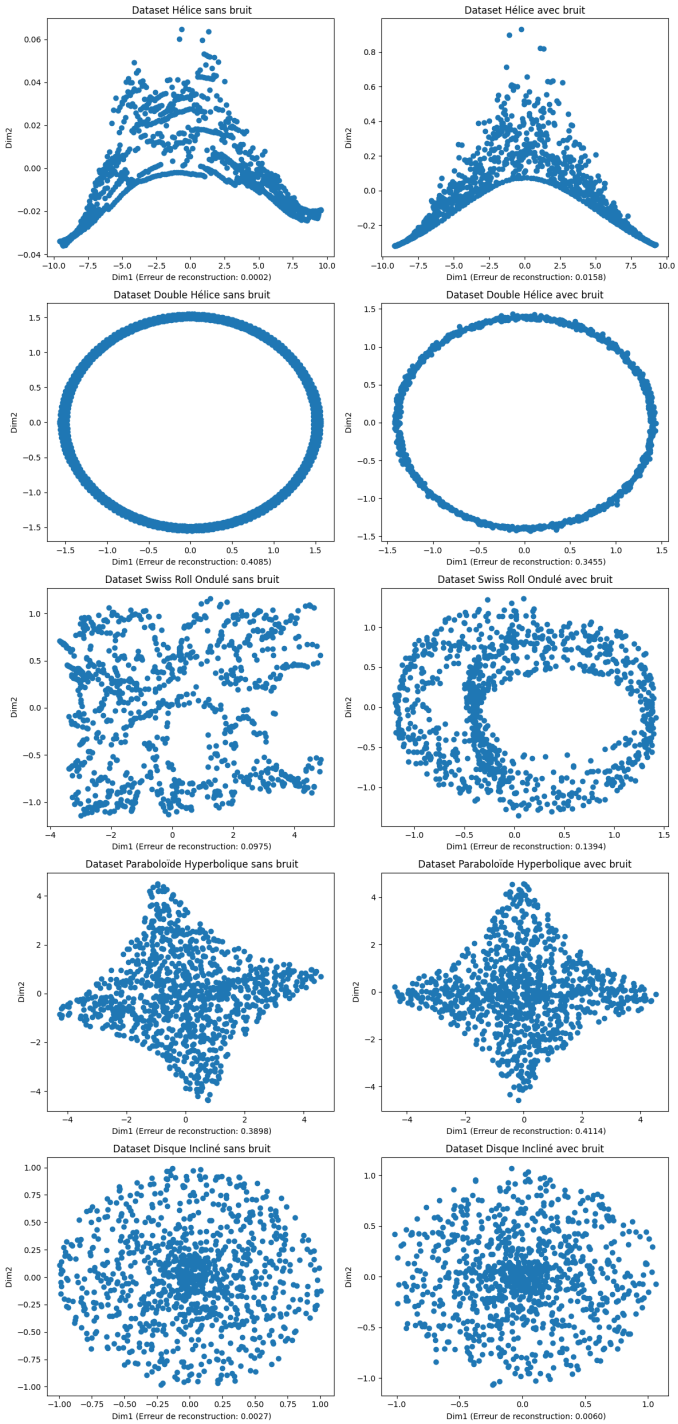

Fig. 9: Résultats de la méthode ISOMAP sur les ensembles de données synthétiques avec et sans bruit.

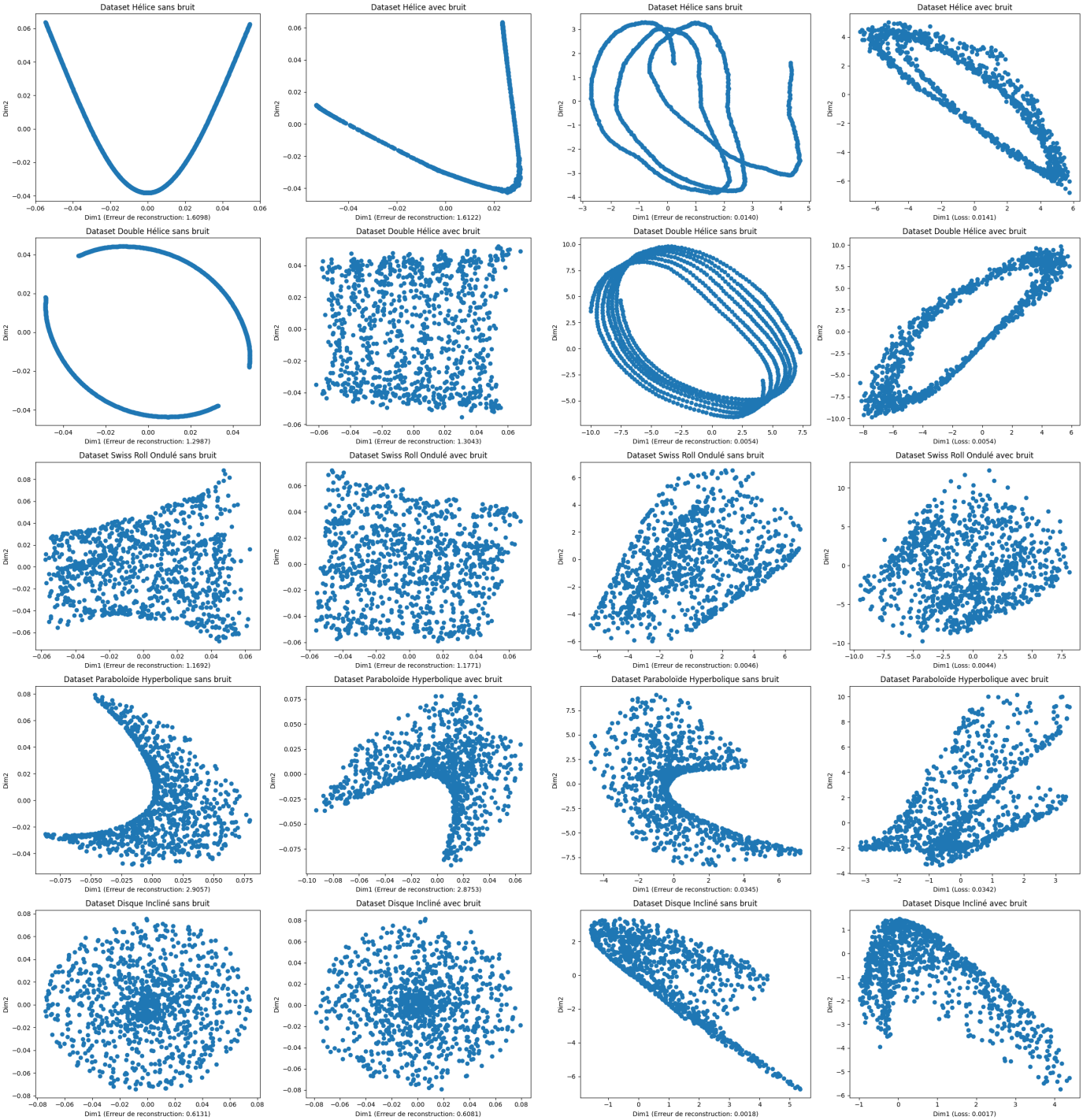

Fig. 10: Résultats de la méthode LLE sur les ensembles de données synthétiques avec et sans bruit.

Fig. 11: Résultats de la méthode Auto-Encoders sur les ensembles de données synthétiques avec et sans bruit.