Octobre 2023

Reduction de dimension : Une étude comparative de l'ACP et de l'Isomap sur des données non-linéaires

Auteur : **BABEY Cassien**
Courriel : cassienbabey@gmail.com

Université Lumière Lyon-2 - Master 2 MALIA

Projet réalisé dans le cadre de l'unité d'enseignement *Manifold learning* Master 2 INFORMATIQUE / MACHINE LEARNING FOR ARTIFICIAL INTELLIGENCE (MALIA).

INFO SUR ARTICLE

Mots clés :

Réduction de dimension

Analyse de Composante Principale (ACP)

Isomap

Trustworthiness

Continuity

1-NN

ABSTRACT

Cette étude évalue les performances de la PCA et de l'Isomap sur des jeux de données artificiels et sur le célèbre jeu de données en haute dimension MNIST. Les méthodes ont été testées sur des structures artificielles telles que le Swiss Roll, l'Helix, et le S-Curve. La comparaison s'appuie sur quatre critères : la visualisation 2D après réduction, la méthode du plus proche voisin (1-NN), la trustworthiness et la continuity. Nos résultats mettent en lumière l'efficacité de chaque technique face à diverses configurations de données et offrent une comparaison basée sur des indices fiables et robustes.

1. Introduction

Le développement de l'industrie et son orientation vers le numérique a entraîné une explosion sans précédent de données multidimensionnelles. Ces données proviennent de diverses sources, allant de la médecine à la finance, en passant par les réseaux sociaux ou encore l'industrie. L'essor de ces données, offrant de nouvelles possibilités en innovation et en recherche, il génère également de nouveaux défis en matière de traitement, d'analyse, de visualisation ou encore d'utilisation. Un des étapes principales dans ce processus est la réduction de dimensionnalité, visant à extraire les informations principales que ces nouvelles données multidimensionnelles nous offrent tout en gardant leur structure

pour mieux les traiter et les utiliser.

Reduction de dimension

La réduction de dimension est une approche qui vise à transformer les données de haute dimension en un ensemble de données de dimension réduite, tout en préservant autant que possible les informations essentielles. Lorsqu'on travaille avec des données de grande dimension, on est souvent confronté au "fléau de la dimensionnalité", un phénomène où la distance entre les points devient quasiment uniforme à mesure que la dimension augmente, rendant les méthodes traditionnelles d'analyse moins efficaces. Ce fléau peut entraîner une détérioration des performances des algorithmes, une augmentation des coûts computationnels, et une difficulté accrue de visualisation. Par conséquent, la réduction de dimension joue un rôle crucial dans l'extraction d'informations significatives, la visualisation, le traitement et la classification des données. Plus que jamais, avec l'augmentation de la complexité et de la dimension des données, cette étape est devenue un élément fondamental des pipelines d'analyse.

Analyse en Composantes Principales (ACP)

L'Analyse en Composantes Principales est l'une des méthodes les plus anciennes et les plus largement adoptées pour la réduction de dimension. La PCA repose sur une transformation linéaire des données originales en un nouvel ensemble de variables appelées composantes principales. Ces composantes sont orthogonales entre elles et captent une quantité décroissante de la variance présente dans les données. Mathématiquement, soit X une matrice centrée de données. La matrice de covariance C est donnée par :

$$C = \frac{1}{n} X^T X$$

où n est le nombre d'échantillons. Les données transformées sont obtenues en projetant X sur les vecteurs propres associés aux plus grandes valeurs propres de C.

Isomap

L'Isomap est une méthode non linéaire qui cherche à conserver les distances géodésiques entre les points dans l'espace de caractéristiques d'origine lors de leur projection dans un espace de dimension inférieure. Le point de départ de l'Isomap est le calcul des distances entre les points. Après avoir défini le graphe des k-plus proches voisins, la distance géodésique entre deux points est la plus courte distance entre eux sur ce graphe. Mathématiquement, si $\boldsymbol{d(i, j)}$ est la distance euclidienne entre deux points $\boldsymbol{i}$ et $\boldsymbol{j}$, et $\boldsymbol{D}$ est la matrice de distances entre tous les points, alors les distances géodésiques $\boldsymbol{G}$ sont données par $\boldsymbol{G} = \boldsymbol{D}$ pour les k-plus proches voisins et $\boldsymbol{G} = \infty$ sinon. Ces distances géodésiques sont ensuite utilisées pour une décomposition en valeurs propres, similaire à la PCA, mais en conservant les distances non linéaires. L'avantage principal de l'Isomap est sa capacité à découvrir des structures non linéaires inhérentes aux données, mais il peut être computationnellement plus coûteux que la PCA et sa performance peut être affectée par le choix des hyperparamètres, comme le nombre de voisins k.

Dans cette étude, nous nous proposons d'entreprendre une analyse comparative approfondie de l'ACP et de l'Isomap. Nous évaluerons leur performance sur des jeux de données artificiels, notamment des structures telles que le Swiss Roll, l'Helix et le S-Curve, ainsi que sur le célèbre jeu de données réel MNIST. En utilisant des critères tels que la visualisation 2D de la réduction de dimension des modèles, le 1-NN pour

déterminer les erreurs de généralisation, la trustworthiness et la continuity, notre ambition est de fournir des insights clairs et exhaustifs sur le comportement et l'efficacité de ces techniques dans différents contextes d'analyse de données multidimensionnelles.

2.Expérimentation

2.1. Données

Dans cette partie, nous expliciterons tout d'abord le choix et l'implémentation de nos données artificielles et réelles, puis nous expliciterons le choix de nos métriques ainsi que notre méthodologie expérimentale pour les tests de performance de nos deux modèles et enfin nous présenterons les résultats qui en découlent.

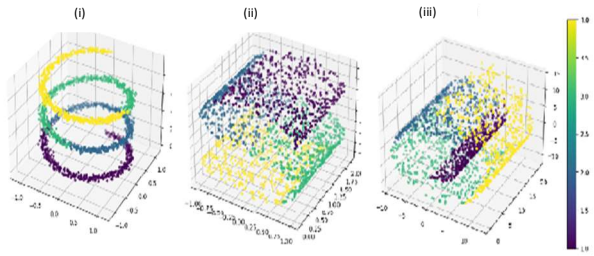
Données artificielles

Dans le cadre de notre expérimentation, nous allons tout d'abord travailler sur des données simulées artificiellement. L'intérêt de travailler sur ces ensembles de données est l'opportunité de contrôler et simuler des caractéristiques et structures spécifiques tout en permettant une reproductibilité de nos expériences. Notre choix s'est penché sur trois structures très connues de données artificielles et souvent utiliser dans la littérature pour observer les performances de modèles de réduction de dimension : l'Helix, la S-Curve et le Swiss Roll. Ces ensembles de données, possédant des niveaux de complexité différents, devraient nous permettre d'étudier efficacement les performances de nos deux modèles dans une tâche de réduction en deux dimensions. Un choix arbitraire de 2 000 données pour chaque set de données a été décidé en accord avec la littérature.

L'Helix se caractérise graphiquement par une structure en forme de spirale ou d'hélice. Au niveau mathématique, sa génération est exprimée à l'aide des fonctions $x(t) = t\cos(t), y(t) = t\sin(t)$ et $z(t) = t$, où t varie sur un intervalle donné. L'intérêt de ce modèle est d'observer si nos deux modèles expérimentaux vont réussir à capturer la structure spéciale de ce type de donnée. Pour la génération de ce jeu de données, nous avons spécialement créé une fonction Python implémentant les fonctions mathématiques de l'Helix.

La S-Curve, quant à elle, se caractérise graphiquement par une structure en forme de « S ». Elle est définie mathématiquement par les fonctions $x(t) = sin(t)$, $y(t) = t$ et $z(t) = cos(t)$, avec t se déplaçant sur un intervalle donné. Nous avons considéré ce jeu de données pour l'observation de la capacité des modèles à « aplatir » sa structure. Pour notre expérimentation, dans un souci de reproductibilité, nous utiliserons la fonction *make_s_curve* de scikit-learn pour générer ce jeu de données.

Enfin, le Swiss Roll se caractérise graphiquement par une structure rappelant une feuille qui s'enroule sur elle-même. Il est conçu en échantillonnant des points d'une feuille rectangulaire, généralement définie par $\boldsymbol{u}$ et $\boldsymbol{v}$, puis en la roulant pour former une spirale à l'aide des transformations $x = u\cos(v)$, $y = u\sin(v)$, et $z = v$. Ce jeu de données teste la capacité d'une méthode à "dérouler" cette structure en spirale tout en conservant les distances locales entre les points. Pour notre expérimentation, nous utiliserons la fonction *make_swiss_roll* de scikit-learn, basé sur l'algorithme de S. Marsland (2014), pour les mêmes raisons qu'évoquées précédemment. Les représentations graphiques de ces trois jeux de données sont représentées dans le graphique 1.



Graphique 1 : Représentation graphique des différents jeux de données artificiels (i. Helix, ii. S-Curve, iii. Swiss Roll)

Comme explicité plus tôt, ces structures tridimensionnelles ont été choisies en raison de leur complexité intrinsèque. Elles mettent en évidence des relations non linéaires entre les points, représentant ainsi un défi pour les méthodes de réduction de dimension. En explorant ces ensembles de données, notre objectif est d'évaluer et de comparer la manière dont nos différentes méthodes parviennent à conserver les relations structurelles entre les données lors du processus de réduction. De surcroît, pour renforcer la pertinence de notre étude, des étiquettes ont été générées pour chacun de ces ensembles en se basant sur la position des points dans l'espace tridimensionnel initial. Ces étiquettes seront essentielles pour évaluer la qualité de la réduction de dimension.

Données réelles

L'objectif de notre étude est également de comparer nos modèles de réduction de dimension sur des données en haute dimension. Pour ce faire, nous avons choisi le jeu de données MNIST (Deng, 2012). Le MNIST (Modified National Institute of Standards and Technology) est l'un des ensembles de données les plus emblématiques dans le domaine de l'apprentissage automatique. Il est constitué d'un ensemble de 70 000 images en niveaux de gris de chiffres manuscrits, allant de 0 à 9. Chaque image a une résolution de 28x28 pixels. La base de données est divisée en 60 000 images pour l'entraînement et 10 000 images pour les tests. La version du MNIST que nous utiliserons est la première version composée de seulement 60 000 images. Nous nous sommes tournés vers ce jeu de données car celui-ci est open source, très bien documenté et prétraité, possède une grande complexité et une grande dimensionnalité. Ces différents aspects vont nous permettre de tester comment nos modèles de réduction de dimension vont se comporter au niveau de la préservation des caractéristiques intrinsèques des données. Enfin, ce jeu de données, étant largement utilisé dans le domaine de la recherche, peut nous permettre de comparer nos résultats avec ceux obtenus par d'autres chercheurs du domaine. Pour notre expérimentation, nous utiliserons un set de données de 5 000 images aléatoirement sélectionnées issues du MNIST. Nous utiliserons la méthodologie employée par Marcus Rottschäfer (2018) dans son projet Github pour importer le set de données.

2.2. Mesure de performance

Pour évaluer de manière optimale l'efficacité de nos méthodes de réduction de dimension, nous avons adopté une méthode multi-facette, similaire à l'article de Van der Maaten et al., 2009, couplant les observations graphiques et quantitatives des réductions effectuées.

Mesure graphique

La visualisation graphique est un outil puissant qui nous permet d'observer directement comment les structures de données de haute dimension sont préservées dans un espace de dimension réduite. L'appréciation visuelle offre un moyen intuitif de discernement des traitements des structures intrinsèques de nos données. Dans le cadre de notre étude, le rendu graphique de la réduction offre un aperçu initial des performances des modèles, permettant d'observer le comportement de nos modèles face à ces jeux de données artificiels.

Trustworthiness

La trustworthiness est une métrique qui quantifie à quel point les voisinages sont préservés lors du passage de l'espace original à l'espace réduit. Mathématiquement, pour un point donné, elle mesure la proportion de points qui, bien qu'étant des voisins dans l'espace réduit, ne l'étaient pas dans l'espace d'origine et est définie par :

$$T(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in U_i^{(k)}} (r(i,j) - k),$$

où *r(j, x_i)* est le rang de *j* parmi les voisins de *x_i* dans l'espace original. Plus cette valeur est élevée, mieux la méthode de réduction de dimension a préservé les relations locales. Cette métrique est particulièrement pertinente pour notre étude car elle offre une indication sur la fiabilité des clusters de points formés post-réduction. Dans notre expérimentation, nous utiliserons la librairie python *coranking* de Lee, John A. et Michel Verleysen (2009) pour calculer cette métrique avec un k de 12, nous permettant une éventuelle comparaison avec l'article de Van der Maaten et al., 2009. Cela nous offre une reproductibilité simplifiée pour d'éventuelles réplications.

Continuity

Inverse de la trustworthiness, la continuity évalue si les points voisins dans l'espace d'origine restent voisins dans l'espace réduit. Mathématiquement, elle mesure la proportion de points qui étaient des voisins dans l'espace d'origine mais ne le sont plus dans l'espace réduit et est définie par :

$$C(k) = 1 - \frac{2}{nk(2n - 3k - 1)} \sum_{i=1}^{n} \sum_{j \in V_i^{(k)}} (\hat{r}(i,j) - k),$$

où *r'(j, x'_i)* est le rang de *j* parmi les voisins de *x'_i* dans l'espace réduit. Une valeur élevée indique que la technique de réduction a conservé efficacement les relations de proximité. Dans notre contexte, cela garantit que les structures intrinsèques des données sont maintenues. Comme pour la trustworthiness, nous utiliserons la librairie python *coraking* pour calculer la continuity et un k de 12.

L'utilisation conjointe de ces différentes métriques pour notre étude comparative devrait nous permettre de fournir une évaluation robuste et complète de nos méthodes de réduction de dimension.

2.3.Algorithmes

Dans cette partie, nous expliciterons la méthodologie que nous avons employé pour réaliser l'expérimentation de nos modèles sur nos différents jeux de données.

Algorithmes pour les données artificielles

Notre méthodologie pour tester les modèles sur les jeux de données artificiels débute par le chargement des données, suivies de leur discrétisation en classes distinctes pour le 1-NN. Nous appliquons ensuite l'Analyse de Composante Principale (ACP) pour réduire la dimensionnalité des données à 2 dimensions. Parallèlement, Isomap est optimisé via une recherche sur grille (Manual GridSearch) pour déterminer le nombre de voisins le plus approprié (dans l'intervalle [5, 10] avec un pas de 5) pour une réduction à 2 dimensions. Les performances de ces techniques sont évaluées à l'aide des erreurs de généralisation 1-NN ainsi que des métriques de trustworthiness et de continuity. Les résultats, consolidés dans un DataFrame, combinent analyses quantitatives et représentations graphiques pour une évaluation complète des techniques de réduction de dimensionnalité. Le pseudo-code suivant explicite la

structure de notre algorithme : *Algorithme pour les données réelles*

Algorithm 1: Évaluation de PCA et Isomap sur des Données Artificielles

Result: DataFrame avec Modèle, Erreur de Classification, Trustworthiness, et Continuity
Charger le jeu de données dans X, y;
Discrétiser les étiquettes en fonction de la troisième colonne de X pour obtenir y;
Trier X selon la troisième colonne;
begin
  Appliquer PCA à X pour obtenir X_pca;
  Calculer l'erreur de généralisation 1-NN pour X_pca pour obtenir pca_erreur;
  Optimiser Isomap (via le score 1-NN) à l'aide de GridSearch manuel pour déterminer le meilleur n_neighbors;
  Appliquer le meilleur modèle Isomap à X pour obtenir X_isomap;
  Calculer l'erreur de généralisation 1-NN pour X_isomap pour obtenir isomap_erreur;
  Calculer trustworthiness et continuity pour X_pca et X_isomap;
  Visualiser les données originales, X_pca, et X_isomap;
  Sauvegarder les résultats dans un dataframe avec les colonnes "Modèle", "Erreur de Classification", "Trustworthiness", et "Continuity";
end

L'algorithme que nous allons utiliser pour nos données réelles est très similaire à notre algorithme pour les données artificielles. La seule différence réside dans la suppression de la visualisation des réductions car nous allons ici passer de données de grandes dimensions à un réduction à 20 dimensions, toujours dans une optique de comparaison avec l'article de Van der Maaten et al., 2009. Le pseudo-code ci-dessous explicite la structure de ce deuxième algorithme :

Algorithm 2: Évaluation de PCA et Isomap sur des Données Réelles

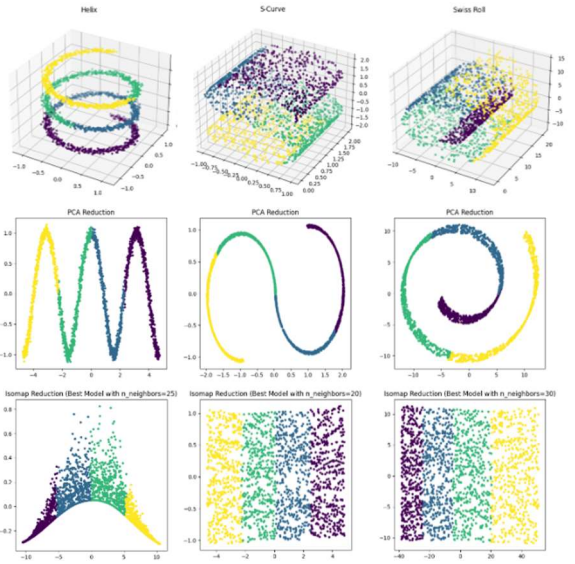
Result: DataFrame avec Modèle, Erreur de Classification, Trustworthiness, et Continuity
Charger le jeu de données dans X, y;
Discrétiser les étiquettes en fonction de la troisième colonne de X pour obtenir y;
Trier X selon la troisième colonne;
begin 2
  Appliquer PCA à X pour obtenir X_pca;
  Calculer l'erreur de généralisation 1-NN pour X_pca pour obtenir pca_erreur;
  Optimiser Isomap (via le score 1-NN) à l'aide de GridSearch manuel pour déterminer le meilleur n_neighbors;
  Appliquer le meilleur modèle Isomap à X pour obtenir X_isomap;
  Calculer l'erreur de généralisation 1-NN pour X_isomap pour obtenir isomap_erreur;
  Calculer trustworthiness et continuity pour X_pca et X_isomap;
  Sauvegarder les résultats dans un dataframe avec les colonnes "Modèle", "Erreur de Classification", "Trustworthiness", et "Continuity";
end

2.4.Résultats

Dans cette partie, nous détaillerons les résultats obtenus par nos algorithmes sur les données artificielles et réelles.

Données artificielles

Le graphique 2 représente les visualisations des réductions des modèles ACP et Isomap pour chacun des sets de données artificiels testés.



Graphique 2 : Représentation graphique des réductions de l'ACPet de l'Isomap pour l'Helix, la S-Curve et le Swiss Roll.

Pour l'ACP, la visualisation semble montrer une compression des données où les spirales sont partiellement superposées par endroit. L'ACP semble réussir à conserver la forme en « S » de la S-Curve mais semble le faire par une compression des données, perdant l'idée d'une conservation réaliste de la structure des données sur un plan intrinsèque. Enfin, au niveau du Swiss Roll, on observe que l'ACP présente une visualisation compressée des données en gardant tout de même la forme en « spirale » de set de données. L'ACP n'a donc pas « déroulé » la structure mais plutôt « compressé » les données pour essayer de reconstruire la structure des données. La génération de nos données a pu influencer ces résultats.

En ce qui concerne l'Isomap, la visualisation obtenue pour l'Helix montre que le modèle a tenté de dérouler la structure en maintenant la distance géodésiques résultat en une visualisation des données présentant une torsion des données, indiquant un déroulement incomplet. Au niveau de la S-Curve et du Swiss Roll, Isomap a parfaitement déroulé les données. Dans l'ensemble Isomap a mieux réussi à « dérouler » les différents sets de données que l'ACP.

Nous avons également rassemblé les scores de performance à notre métrique d'erreurs de généralisation sur nos jeux de données artificiels dans le tableau 1. Une valeur faible correspond à un bon score pour cette métrique.

Dataset	*ACP*	*Isomap*
Helix	14.85%	**14.50%**
S-Curve	14.9%	**14.7%**
Swiss Roll	15.4%	**14.6%**

Tableau 1 : Erreur de généralisation (en %) du 1-NN Classifiers entrainés sur les données artificielles

Pour l'ensemble Helix, l'ACP a résulté en un taux d'erreur de 14.85% contre **14.50%** l'Isomap. Pour la S-Curve, les taux d'erreur étaient respectivement de 14.9% avec l'ACP et de **14.7%** avec l'Isomap. Enfin, avec le Swiss Roll, le taux d'erreur obtenue avec l'ACP était de 15.4%, tandis qu'Isomap a permis de réduire cette erreur à **14.6%**. Ces résultats indiquent que, pour ces jeux de données, Isomap a systématiquement surpassé l'ACP en termes d'erreur de généralisation du classificateur 1-NN.

Au niveau des scores de trustworthiness, pour l'ensemble Helix, l'ACP a produit un score de 0,997685 contre **0.998759** pourl'Isomap. Pour la S-Curve, l'ACP a donné un score de 0.936938, alors qu'Isomap a presque atteint un score parfait avec un score de **0.999904**. Enfin, avec le Swiss Roll, l'ACP a obtenu un score de 0,950076 contre un score encore proche de la perfection pour Isomap (**0.999957**). Ces résultats montrent qu'Isomap obtient encore des scores de trustworthiness supérieurs à l'ACP pour nos sets de données artificiels. Les résultats sont rapportés dans le Tableau 2.

Dataset	*ACP*	*Isomap*
Helix	0.997685	**0.998759**
S-Curve	0.936938	**0.999904**
Swiss Roll	0.950076	**0.999957**

Tableau 2 : Trustworthiness T(12) sur les données artificielles.

Les derniers scores de métrique a étudié est la « continuity ». Les scores obtenus lors de notre expérimentation sont rapportés dans le Tableau 3.

Dataset	*ACP*	*Isomap*
Helix	**0.998797**	0.998610
S-Curve	0.987139	**0.999894**
Swiss Roll	0.987752	**0.999953**

Tableau 3 : Continuity C(12) sur les données artificielles.

Pour l'ensemble Helix, l'ACP a produit un score de **0,998797**, tandis qu'Isomap a affiché un score légèrement inférieur de 0,998610. Pour la S-Curve, l'ACP a enregistré un score de 0,987139, alors qu'Isomap a atteint un score remarquablement élevé de **0,999894**. Enfin, avec le Swiss Roll, l'ACP a obtenu un score de 0,987752contre un score de **0.999953** pour l'Isomap. Il est à noter que, bien qu'Isomap ait surpassé l'ACP en termes de "Continuity" pour la S-Curve et le Swiss Roll, l'ACP a légèrement devancé Isomap pour le jeu de données Helix.

Données réelles

Le tableau 4 présente une comparaison des scores obtenus pour les modèles ACP et Isomap sur un jeu de données réelles. Pour l'erreur de généralisation, l'ACP a enregistré un score de 0,5392, alors qu'Isomap a affiché un score légèrement inférieur, plus performant, de **0,5046**. En ce qui concerne la Trustworthiness, l'ACP a démontré une excellente performance avec un score de **0,997601**, largement supérieur à celui d'Isomap qui est de 0,759299. Enfin, pour la Continuity, l'ACP a également surpassé Isomap avec un score de **0,998776** par rapport à 0,939698 pour Isomap. Ces résultats montrent une supériorité notable de l'ACP en termes de Trustworthiness et de Continuity sur les données MNIST, bien qu'Isomap ait montré une performance

légèrement supérieure en termes d'erreur de généralisation.

Modèle	*1-NN*	*T(12)*	*C(12)*
ACP	53.92%	**0.997601**	**0.998776**
Isomap	**50.46%**	0.759299	0.939698

Tableau 4 : Score d'erreur de généralisation (*1-NN*), de Trusworthiness (*T(12)*) et de Continuity (*C(12)*) sur les données réelles.

3. Discussion

L'objectif de notre étude est de proposer une comparaison robuste et complète des performances de l'ACP et d'Isomap dans une tâche de réduction de dimension. Nous avons tout d'abord orienté notre étude vers une comparaison sur des données artificielles simulées. Les résultats obtenus sur ces sets de données nous permettent de dégager certaines informations importantes pour notre étude. Les scores à nos différentes métriques quantitatives nous montrent que les deux modèles affichent des performances relativement élevées d'un part, et très proche de l'autre. En se basant sur la théorie générale que l'ACP performe mal sur des données non-linéaires, nous nous attendions à des performances plus encartées entre nos deux modèles or nos observations montrent, qu'au niveau de nos métriques quantitative du moins, les deux modèles proposent des performances proches et élevées. Ces résultats corroborent les résultats de Van der Maaten et al., (2009) qui ont obtenu des scores de Trustworthiness et Continuity très élevé pour les deux modèles sur des jeux de données similaire dans leur étude. La seule différence notable est la meilleure performance de notre ACP sur la métrique d'erreur de généralisation. Cependant, cela peut s'expliquer par nos méthodes de génération de données ou de calcul de métrique qui diffèrent de celles de l'article.

Notre étude apporte également des éclaircissements supplémentaires sur les capacités et limites des deux méthodes sur le maintien de la structure des données lors de la réduction de dimension par l'apport de notre visualisation. Nous pouvons observer que l'ACP, qui est une technique linéaire, offre des représentations que l'on pourrait décrire comme « simplifiées » des structures des modèles. Cependant, cette simplification implique une potentielle perte d'informations des données. La méthode de l'ACP ne semble donc pas être une méthode optimale pour traiter ce type de données. A l'inverse, Isomap, qui est une technique non-linéaire, a montré des résultats très convaincant pour les jeux de données Swiss Roll et S-Curve en « déroulant » presque parfaitement les deux jeux de données. Toutefois, il est à noter que le modèle peine à produire une représentation optimale du set de données Helix, suggérant que certaines structures peuvent présenter des défis même pour des méthodes non-linéaires. Prendre en compte la visualisation des réductions des méthodes nous a permis de mettre en exergue l'importance de bien choisir la méthode adaptée selon nos données.

Toujours pour notre étude, nous avons tester nos deux modèles sur des données réelles issues du set de données MNIST afin d'observer les performances de ces modèles dans des conditions réelles. Les résultats obtenus nous ont permis d'apporter une nuance importante sur nos interprétations. En effet, l'ACP a montré une robustesse impressionnante, indiquant que le modèle arrive à maintenir fidèlement les proximités entre les échantillons voisins et à respecter l'ordre global des données, suggérant une capacité à saisir les tendances linéaires dominantes dans les données réelles. A l'inverse, le modèle Isomap, qui présentait des performances supérieures sur les données artificielles, semblent être

moins robuste que l'ACP sur des données réelles. Cependant, le modèle présente une meilleure performance sur le taux d'erreur de généralisation, indiquant qu'il arrive à mieux préserver les structures et relations intrinsèques des données, préservation qui pourrait être omise ou simplifiée par une approche purement linéaire comme l'ACP. Il est tout de même important de noter que le taux d'erreur de généralisation augmente drastiquement entre les données artificielles et les données réelles pour les deux modèles. Ces résultats corroborent également les résultats obtenus par Van der Maaten et al., selon lesquels l'ACP proposerait de meilleures performances sur ce set de données MNIST.

En conclusion, notre étude a permis de mettre en avant le comportement des modèles d'ACP et d'Isomap dans un contexte de simulation et dans un contexte de mise en application réelle. Nous avons pu observer les forces et faiblesses de ces deux modèles dans ces différents contexte expérimentaux, mettant en exergue l'importance de ne pas toujours se baser sur les scores de métriques d'une part, et de bien prendre le contexte d'autre part. Dans certains contextes, la combinaison de ces deux méthodes pourrait s'avérer être bénéfique pour exploiter leurs forces respectives. En fin de compte, cette étude souligne l'importance de la prudence et de la diligence lors de la sélection d'outils pour l'analyse de données, en gardant à l'esprit les avantages et les limites de chaque méthode.

FIN

Données MNIST

Les données MNIST utilisées pour cette étude sont issues du site Kaggle : https://www.kaggle.com/datasets/hojjatk/mnist-dataset?select=train-labels.idx1-ubyte

Fonctions Python utilisées :

La fonction d'importation des données est issue du Github : https://github.com/shukali/dimensionality-reduction-comparison/tree/master

La fonction 1-NN est disponible sur le Github : https://github.com/MaayanLab/Graph-DR/tree/master

Documentation package coranking:

La documentation pour le package Python coranking est disponible ici : https://coranking.readthedocs.io/en/latest/#

Références

- Zhang, Y., Shang, Q., & Zhang, G. (2021). pyDRMetrics - A Python toolkit for dimensionality reduction quality assessment. *Heliyon*, *7*(2), e06199. https://doi.org/10.1016/j.heliyon.2021.e06199

- Das, Suchismita & Pal, Nikhil. (2020). Nonlinear Dimensionality Reduction for Data Visualization: An Unsupervised Fuzzy Rule-based Approach.

- van der Maaten, L., Postma E., & van den Herik J. (2009). ***Dimensionality Reduction: A Comparative Review.***

- S. Marsland, "Machine Learning: An Algorithmic Perspective", 2nd edition, Chapter 6, 2014. https://homepages.ecs.vuw.ac.nz/~marslast/Code/Ch6/lle.py

- Lee, John A., and Michel Verleysen. "Quality assessment of dimensionality reduction: Rank-based criteria." Neurocomputing 72.7 (2009): 1431-1443.