Exploration visuelle des données

Mohamed BEN HAMDOUNE[1] and Lucas ISCOVICI[1]

[1]Université Paris-Descartes, 12 Rue de l'École de Médecine, 75006 Paris, France

Abstract

L'interprétation et l'exploration visuelle des données des jeu de données multidimensionnels peut être un véritable casse-tête pour l'humain. Tandis que les jeux de données en 2D/3D sont très facilement affichables: il est aisé de représenter la structure inhérente de leurs données, il n'est à l'inverse pas facile de réaliser une visualisation intuitive pour des jeux de données à grande dimension. Afin d'améliorer la capacité de visualisation et d'interprétation de ces grands jeu de données multidimensionnels, la dimension doit être réduite, tout en essayant de garder le maximum d'informations. L'objectif des méthodes présentées dans ce papier sont doubles: réaliser une réduction de la dimension par extraction de caractéristiques mais également permettre l'interprétation visuelle des informations conservées

1 Introduction

Nous allons étudier les techniques de visualisation vues en cours ainsi que les techniques de réduction de dimension permettant d'améliorer l'apprentissage et son interprétation. Il convient de rappeler que la famille des méthodes d'analyse des données peut se subdiviser en deux familles de méthodes:

1. Linéaire: les axes utilisés sont des combinaisons linéaires des variables initiales. Ces algorithmes essaient de choisir la projection linéaire "optimale/intéréssante" des points sur un espace vectoriel réduit. Elles peuvent s'avérer puissantes, mais ne prennent pas en compte les structures non linéaires dans les données.

2. Non-linéaires: La plupart de ces méthodes se basent sur la notion de voisinage, avec des graphes, ou simplement des combinaison linéaire de voisins. Ces méthodes s'avèrent plus sensibles aux structures de données non linéaires et essaye de préserver les propriétés globale/locales des données.

Ces deux types de méthodes se basent tout deux sur l'idée que la dimension de beaucoup de jeux de données est souvent artificiellement élevée et qu'il est possible de réduire celle-ci sans relever une perte notable d'information.
Dans un premier temps nous présenterons de manière assez brève chacun des algorithmes utilisés, puis nous réaliserons une étude comparative de certains algorithmes de réduction de dimension linéaire et non-linéaires sur des jeux de données : Gordon et Pomeroy.

2 Algorithmes de réduction de dimension par extraction de caractéristiques

2.1 Approche Linéaire

2.1.1 Analyse en Composantes Principales

L'ACP recherche un espace permettant de rendre compte de la forme du nuage initiales en minimisant les deformations de la projection.
Elle va simplement utiliser la matrice de variance-covariance en la diagonalisant pour en extraire les valeurs et vecteurs propres. Les composantes principales seront construites à partir de des valeurs/vecteurs propres. Ainsi elles seront orthogonales et donc décorées les unes de autres. Ces axes sont composés uniquement des combinaisons linéaires des variables initiales. Cette méthode est une des première méthode d'analyse factorielle et moteur central de toute autre méthode d'analyse factorielle. Dans notre cas, nous projetterons les points du jeux de données sur les premiers axes factoriels (1 et 2) créant ainsi un nouvel espace vectoriel.

2.1.2 Analyse Linéaire Discriminante

L'analyse discriminante linéaire consiste avec la connaissances des classes à trouver un espace à faible dimension optimal telle que, lorsque les points sont projetés, les données de différentes classes sont bien séparée. Cette technique ce base essentiellement sur la matrice de covariance intra-classes et la matrice de covariance inter-classes.
La LDA en tant que classificateur est une solution rapide avec d'excellents résultats, mais si la caractéristique du jeu de données est non linéaire, LDA sera inefficaces. Nous avons vu des travaux où les auteurs combinent PCA et LDA, tous deux pour la réduction de la dimensionnalité. Premièrement, PCA agit en trouvant des composantes principales décorellé, puis la LDA est appliqué en réduction de dimension linéaire. [1].

2.1.3 Multi-Dimentional Scaling

MDS cherche une représentation des données dans un espace vectoriel réduit, pour lequel, la projection des individus respecte bien les distances et cherche à minimiser l'erreur à celle de l'espace vectoriel d'origine de dimension bien superieure. En général, c'est une technique utilisée pour analyser la similarité ou la dissimilarité entre les données. Ces dissimilarités sont représentées sous forme de distances dans un espace géométrique. Une vue simplifiée de l'algorithme est la suivante:

1. Attribuez des points à des coordonnées arbitraires dans un espace à n dimensions défini.
2. Calculez les distances euclidiennes (ou en choisissant une autre distance) parmi toutes les paires de points pour former la matrice.
3. Comparez la matrice en construction avec la matrice d'entrée D en évaluant la fonction de contrainte (la fonction Stress est la fonction coût dans notre cas), plus la valeur est petite et plus la correspondance entre les deux est grande.
4. Répétez les étapes 2 et 3 jusqu'à ce que le *stress* ne diminue plus.

2.2 Approches non-linéaires

2.2.1 Isomap

Isomap peut être considéré comme une extension de MDS ou de K-PCA. Cet algorithme cherche un espace de dimension réduite qui conserve les distances "géodésiques" entre tous les points. Cette méthode se base sur les plus proches voisins de chacune des instances afin de conserver cette distance. Premièrement, Isomap génère un graphe de voisinage entre toutes les instances du jeu de données, puis construit une table de dissimilarité

entre les points grâces à la représentation sous forme de graphe préalablement générée et l'algorithme de Djisktra (ou Floyd) afin de voir quels sont les points les plus proches de chacun. Puis finalement, une fois cette table de dissimilarité créée, on utilise MDS dessus.

2.2.2 Locally Linear Embedding

LLE est une multitude de méthodes factorielles linéaires comparées et concaténées une-à-une dans des zones locales. On pourrait le comparer à une succession d'ACP lancées dans des zones locales du jeu de données puis comparées globalement afin de garder la meilleur qualité des données non-linéaires. Il se base sur une intuition geometrique: on suppose que tous les points du jeux de données ainsi que ces voisins sont séparables ou disposés d'une manière linéaire dans une zone locale. Une variante Hessienne de LLE existe: également connue sous le nom de Hessian Eigenmapping, il semblerait que LLE soit sensible au problème de la régularisation. Quand le nombre de voisins est superieur au nombre de dimensions, la matrice gardant en mémoire chacun des champs locaux linéaires n'a pas suffisament da rangs. Pour éviter ce problème de régularisation, on utilise une forme hessienne pour chaque groupe local afin de bien définir la structure linéaire

3 Jeux de données

3.1 Gordon

Le jeux de données suivant provient d'une recherche [2] et ils ont testé la fidélité du diagnostic basé sur le ratio dans différents tissus (31 MPM et 150 ADCA).
Un ensemble de formation de 32 échantillons (16 MPA et 16 ADCA) a été utilisé pour identifier des homologues de gènes présentant des niveaux d'expression hautement significatifs et inversement corrélés afin de former un total de 15 rapports de diagnostic à l'aide de données de profil d'expression.
Ils ont ensuite examiné (dans le jeu de tests) la précision de plusieurs ratios combinés à un simple outil de diagnostic. En utilisant deux et trois ratios d'expression, Ils ont constaté que les diagnostics différentiels de MPM et d'ADCA pulmonaire étaient précis à 95% et 99%, respectivement. Ils proposent donc l'utilisation de l'expression génique qui est une technique précise et peu coûteuse pouvant être appliquée directement à la clinique pour distinguer le MPM du cancer du poumon.

3.2 Pomeroy

Le jeux de données suivant est aussi un exemple provenant de la médecine [3]. Cet ensemble comporte des données d'expression génétique des micros réseaux d'ADN. Notre jeu de donnée est composé de 42 objets et 1379 variables. Nous avons aussi 5 classes qui sont *MD, Mglio, Rhab, Ncer, PNET*.

4 Analyse

4.1 Poomery

Nous commençons par faire une l'ACP, il est important de noter que les valeurs propres inférieur à 1 indique que la composante principale concernée représente moins de variance par rapport à une seule variable d'origine, lorsque les données sont standardisées. Donc notre cas les 41 premières composantes principales ont des valeurs propre strictement supérieur 1 donc apportent de l'information. Ceci est généralement utilisé comme seuil à partir duquel les composantes sont conservés et cela ne s'applique que lorsque les données sont normalisées.

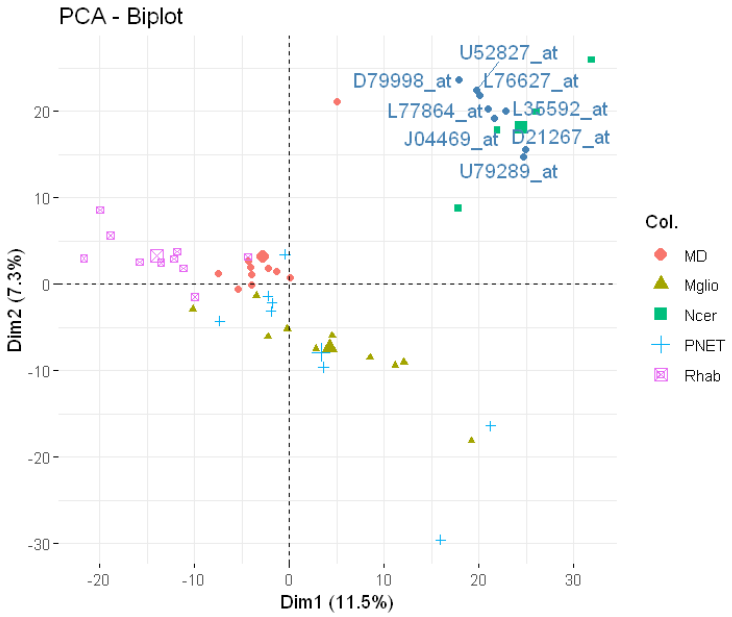


Figure 1: BiPlot ACP Pomeroy

Maintenant nous avons nos 41 principales composantes, c'est une réduction du jeux de donnée de l'ordre de 97% (1379 variables aux départ). Nous avons effectuer une projection sur les deux premiers plan factoriel [13] et constater que la séparation des classes n'est pas capté par l'ACP.
Ensuite, nous avons examiné le cercle de corrélation [14] pour les variables contribuant à hauteur de 80% et plus (8 variables sont présentes). Le plan factoriel contient environ 18% d'information. De plus, sans effectuer une étude approfondi sur les gènes, le première axe est lié à D21267_at et le second axe [15] est lié à *D79998_at, U52827_at, L76627_at.*
La classe *Ncer* est donc fortement caractérisé par *D21267_at* et *D79998_at.*

Nous procédons à une ADL en utilisant les données de l'ACP, le jeux de données en initialisation n'aura aucunes de ses variables linéairement corrélés. La «proportion de trace» est le pourcentage de séparation atteint par chaque fonction discriminante. Pour les données de Poomery, nous obtenons 73,51% 19,17% 7,13% 0,01%.
Notre nombre d'observation est à 42 et notre nombre de variables est 41, nous avons décider de réaliser une ADL avec 34 composantes principales (contient 95% de l'information) afin d'enlever le problème de corrélation (non linéaire) entre les variables.

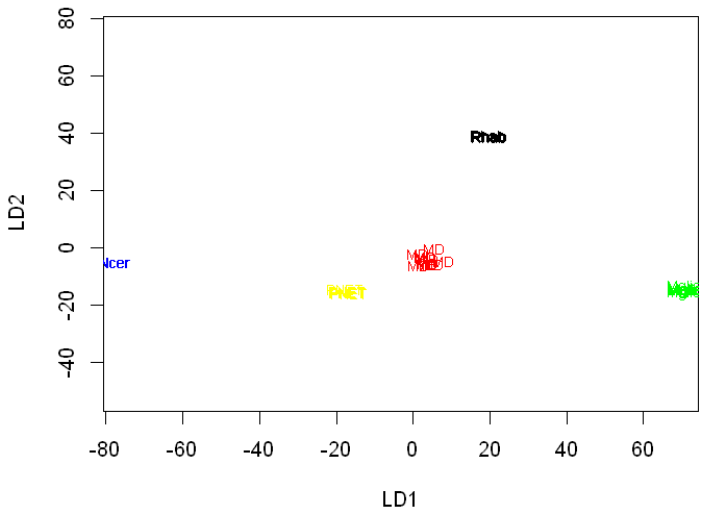


Figure 2: Principal facteur avec ADL

Le premier axe a en effet un grand pouvoir de discrimination, chaque classe projeter sur l'axe *LDA1* est bien séparer de manière ce que la variance intra-classe est minimiser puis la variance inter-classe est maximiser.
Ensuite la second axe ayant une valeur de discrimination relativement plus faible, nous pouvons constater que les variances inter-classes ne sont pas maximiser, sauf pour la classe *Rhab*.

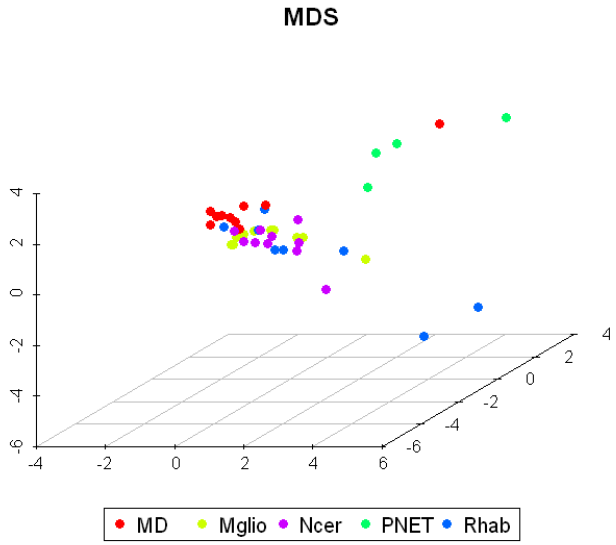


Figure 3: MDS avec une matrice de dissimilarité euclidienne

Nous pouvons voir une certaine séparation des classes dans l'espace à 3 dimensions, mais en revanche

dans notre cas la métrique euclidienne n'est peut-être pas la bonne. En effet le résultat du GOF (Goodness of fit) est moyennement faible 0.24, Nous pensons donc que la réduction de dimension est trop drastique (1379 dimensions à la base) pour cette méthode et de plus la métrique euclidienne pour le problème des gènes n'est peut être pas la plus appropriée.

Nous allons maintenant dérouler les deux méthodes non linéaires vues en cours (Isomap et LLE). La LLE préserve les propriétés locales des données en représentant chaque échantillon dans les données par une combinaison linéaire de ses k plus proches voisins avec chaque voisin pondéré indépendamment. Enfin, LLE choisit la représentation de faible dimension qui préserve le mieux les poids dans l'espace cible. La fonction LLE effectue cette réduction de dimension pour une dimension donner *dim* et voisins *k*.
Nous avons volontairement choisi 41 voisins puisque nous avons 42 observations et une représentation dans un espace de dimension 3.

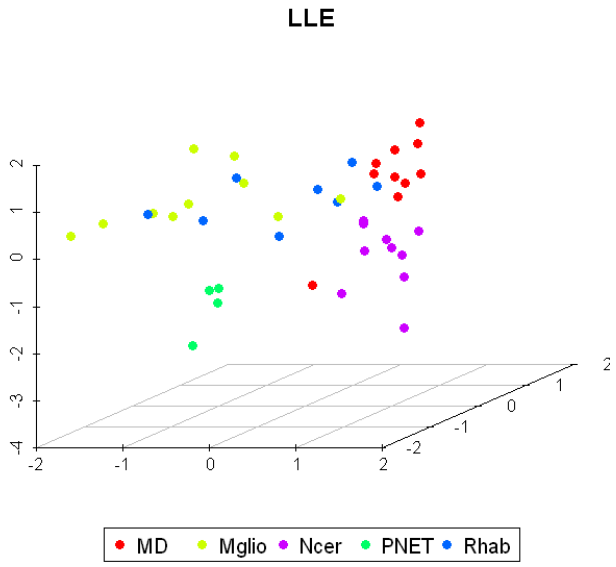


Figure 4: Représentation en 3 dimensions de la LLE

La LLE arrive à faire une séparation des classes plus distinctes que la MDS vu précédemment. Nous avons choisi un nombre éleve de voisins proches car nous avons très peu de d'observation et il en ressort que la classe *MD* n'est plus mélanger comme on a pu le voir sur l'ACP.

En revanche, les résultats de l'Isomap sont plus difficilement interprétable, la séparabailité n'est pas aussi marquante qu'avec la LLE. Contrairement à LLE, elle préserve les propriétés globales des données. Cela veut dire que les distances géodésiques entre tous les échantillons sont mieux capturées lors de l'intégration en basse dimension. Cette implémentation utilise l'algorithme de Floyd pour calculer le graphe de voisinage des distances les plus courtes lors du calcul des distances géodésiques. La fonction Isomap effectue cette réduction de dimension pour un vecteur donné de dimensions dims et les voisins k et cette implémentation comprend en outre une version modifiée de l'algorithme Isomap original, qui respecte voisins les plus proches et les plus éloignés (paramètre *mod* à TRUE).

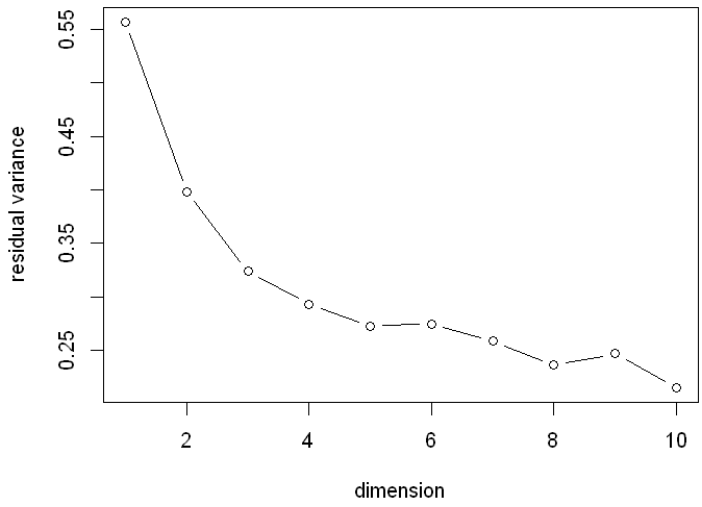


Figure 5: Isomap avec 41 voisins proches

Un coude est formé pour une dimension égale à 3, on projette sur 2 dimension afin de faciliter la visualisation:

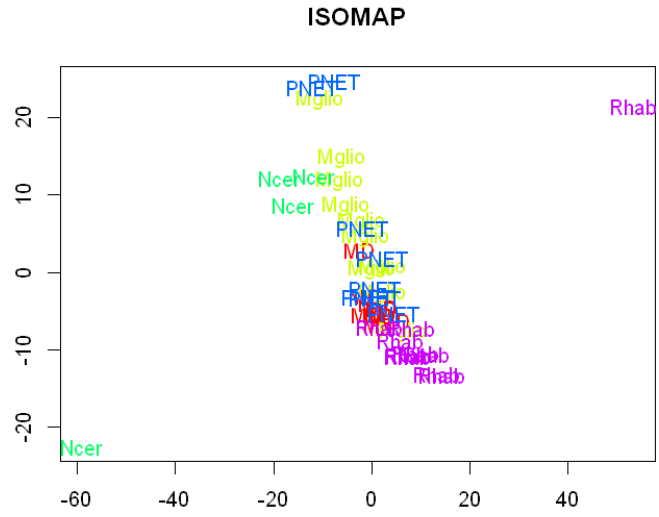


Figure 6: Représentation en 3 dimensions de l'Isomap

La classe *MD* est ici mélanger par rapport aux autres classes, c'est à dire qu'elle ne possède pas de topologie locale propre comme les individus appartenant à la classe *PNET*.
L'Isomap capture l'idée des gènes à travers les différentes classes sans regarder la topologie locale de chaque individus.

4.2 Gordon

Nous sommes dans le cas d'un jeu de données supervisé avec 2 classes. Nous commençons l'analyse avec une ACP, nous avons 95% de l'information contenu dans les 137 premières composantes principales.

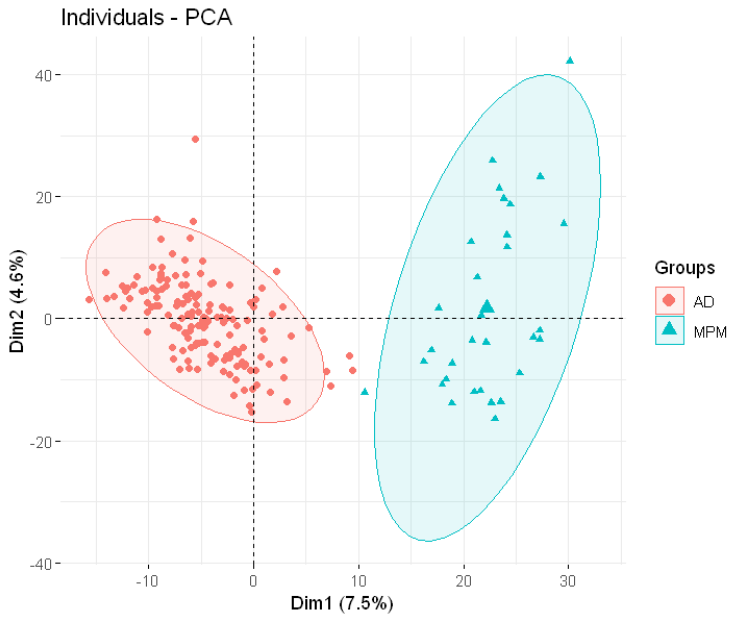
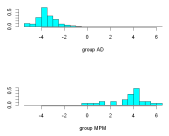
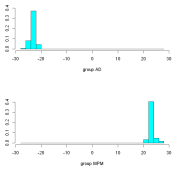


Figure 7: Projection des deux individus sur les deux premiers plan factoriels

À première vu, les données semble facilement être linéairement séparable. les proportions des individus sont 82,87% appartenant à la classe *AD* et ensuite 17,12% à la classe *MPM*. En revanche, nous pouvons noter que seulement 12.1% de l'information ont sur les deux premières composantes principales.
Pour la suite, nous effectuons une ADL avec une initialisation avec ou sans l'ACP afin de montrer la différence flagrante sur la capacité a discriminée les 2 classes.



(a) ADL sans initialisation



(b) ADL avec initialisation

Figure 8: Analyse Discriminante Linéaire sur Gordon

Dans les deux cas, les deux classes sont parfaitement séparées suivant la première composante principale.

Nous continuons avec une approche linéaire, la MDS avec une distance euclidienne sur 2 dimension. Le GOF (Goodness Of Fit) retourné est de 0.12, ce qui reste moyen.

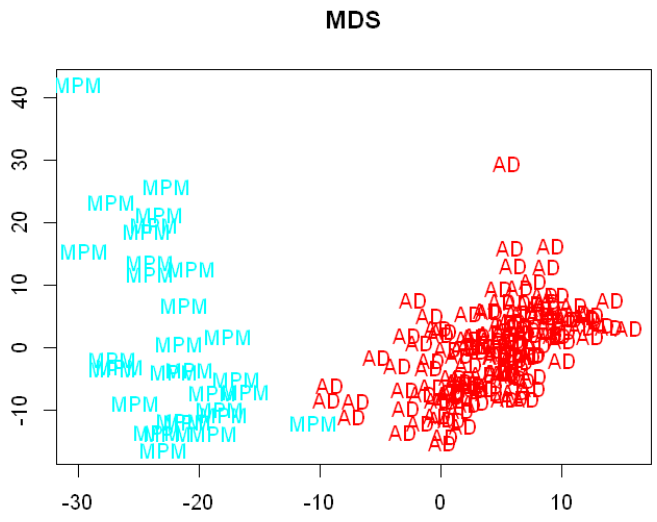


Figure 9: MDS distance euclidienne sur 2 dimension

Nous avons ensuite voulu réaliser une Isomap, en se basant cette fois encore sur le critère du coude pour le choix de la dimension.

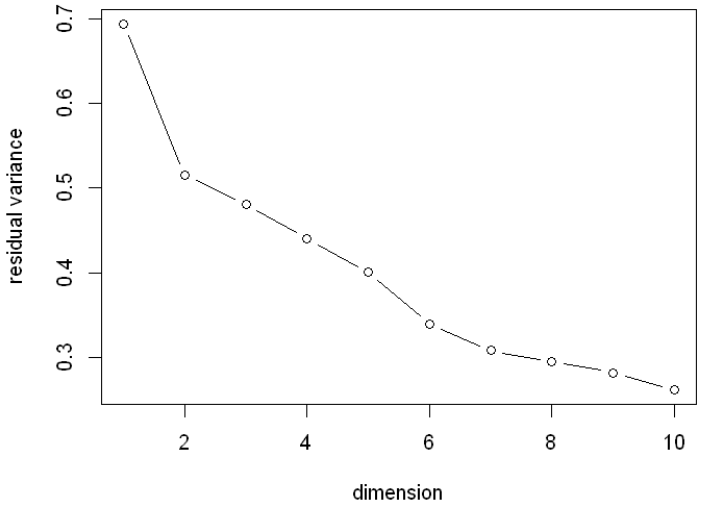


Figure 10: Choix de la dimension pour Isomap

Il apparaît sur ce graphique un coude lorsque la dimension est égale à 2.
En exécutant une Isomap à 2 dimensions et le nombre de proches voisins à 179, nous arrivons a bien séparer les classes. On remarque une grande densité des individus et une bonne discrimination ce qui permettrai de faciliter un modèle d'apprentissage automatique et avoir de bon résultat.

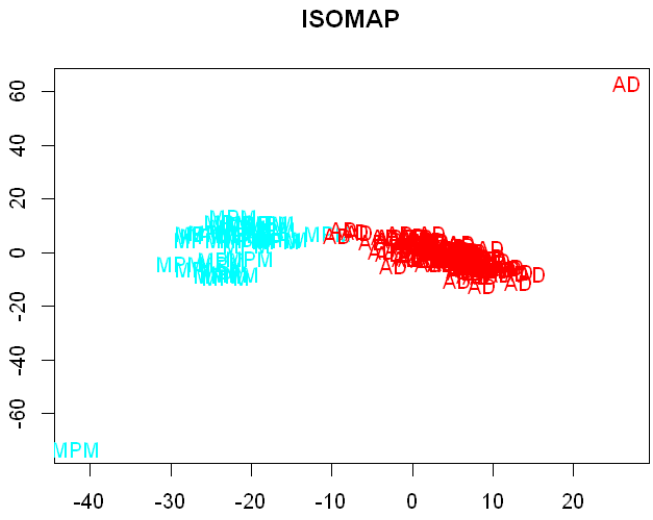


Figure 11: Isomap sur 2 dimensions

Nous finalisons la visualisation avec la LLE :

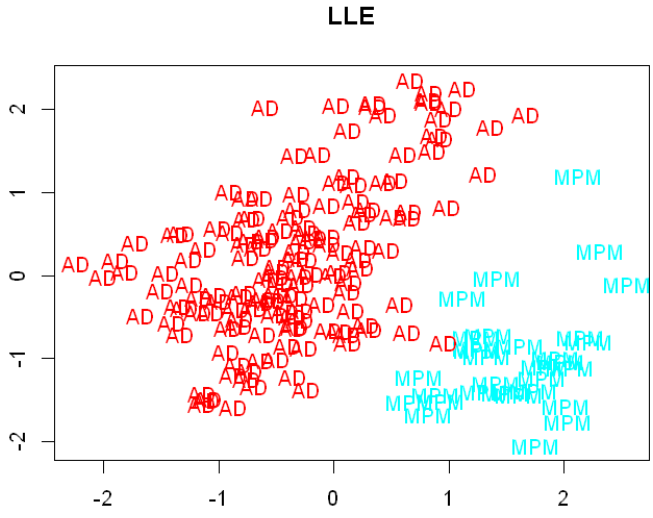


Figure 12: LLE 2 dimensions

Nous remarquons une séparation entre les deux, mais sans une étiquetage du jeu de données il serait plus difficile de réaliser un clustering à partir de cette méthode car on remarque qu'il y a une faible densité sur le graphique et les points sont plus éparpiller. Cela tiens à la méthode qui cherche discriminer localement chaque partie du jeu de données.

5 Conclusion

Nous avons abordé dans le cadre de ce projet différentes techniques utilisées pour la visualisation et l'analyse des données. Ces techniques peuvent servir à la réduction de dimensions de données à plusieurs variables explicatives et à leur classification. Nous avons remarquer l'ACP peux être utile pour décorréler les variables initiale et ainsi être utiliser pour une ADL. On a aussi remarqué que la ADL du fait de ces connaissances des classes, fonctionne à merveille, malgré qu'elle ne puisse pas capter les phénomènes non-linéaire.
L'ACP utilisé comme initialisation pour l'ADL facilite l'interprétation et les améliore les résultats dans certains cas.
l'ACP cherche des relations entre les variables alors que la MDS cherche des similarités entre les observations ce qui fait qu'on peut interpréter les axes de l'ACP à l'aide des variables alors que la MDS ne le permet pas (tout simplement parce qu'il n'y a plus de variables). La MDS reste donc plus une méthode de réduction de dimension que de visualisation.
Pour la LLE et l'Isomap, nous avons fait varier le nombre de voisins aux maximum proches pour chaque jeux de données car il reste extrêmement faible. L'Isomap a tendance à avoir un nuage de points où les classes sont séparées et dense alors que la LLE aura tendance à chercher localement chaque particularité des individus et donne un éparpillement lors de la visualisation de chaque individus tout en restant lisible.
Finalement, ce projet nous a permis de réaliser plusieurs techniques vues en cours sur deux jeux de données où le nombre d'observation d'observation étaient extrêmement faible par rapport aux variables.

A Annexe

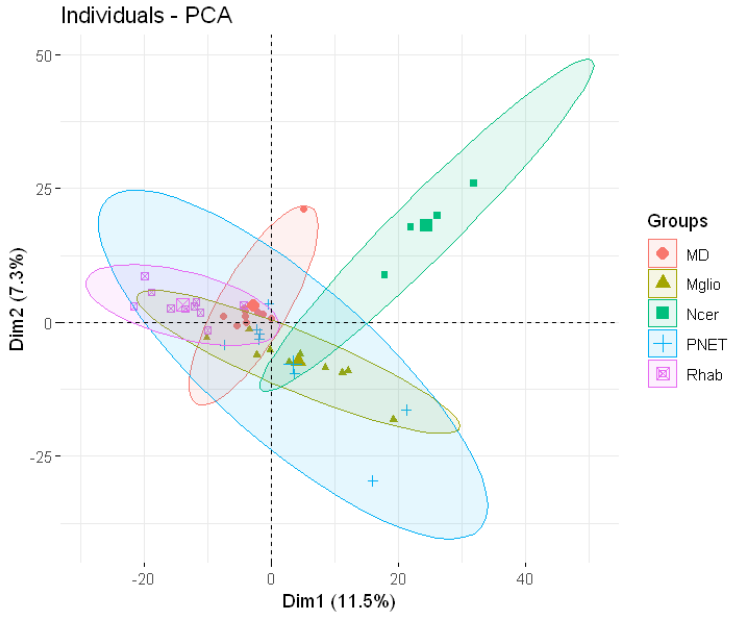


Figure 13: Projection des individus et leurs classes respectifs

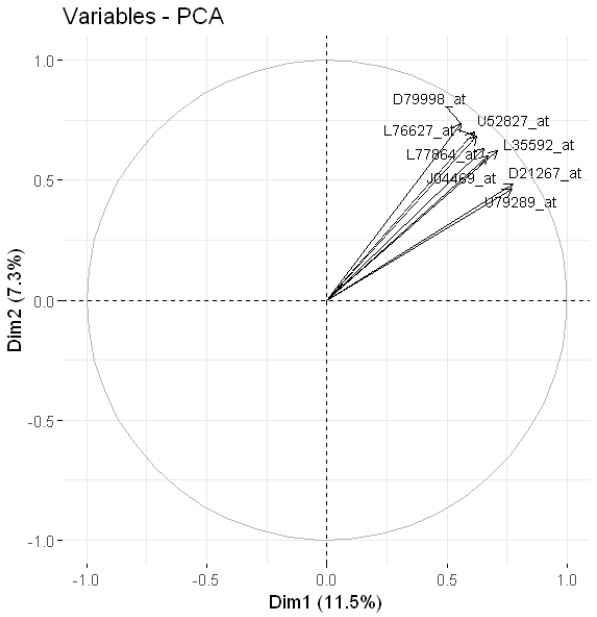


Figure 14: Cercle de corrélation pour Poomery

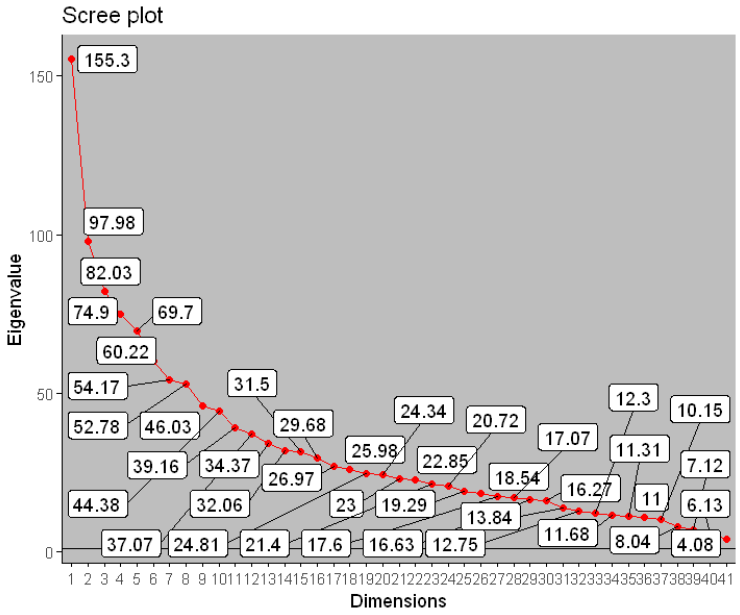


Figure 15: Les valeurs propres de chaque dimension

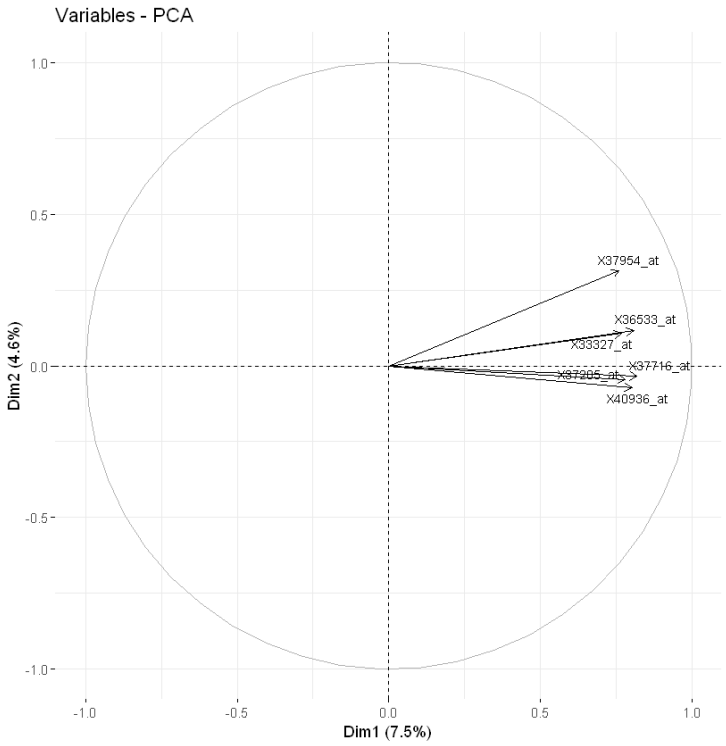


Figure 16: Cercle de corrélation pour Gordon

References

[1] H. Su and X. Wang. Principal component analysis in linear discriminant analysis space for face recognition. pages 30–34, Nov 2014.

[2] Gavin J Gordon, Roderick Jensen, Li-Li Hsiao, Steven R Gullans, Joshua Blumenstock, Sridhar Ramaswamy, William G Richards, David Sugarbaker, and Raphael Bueno. Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma. *Cancer research*, 62:4963–7, 10 2002.

[3] Scott Pomeroy, Pablo Tamayo, Michelle Gaasenbeek, L-M Sturla, Michael Angelo, Margaret E McLaughlin, John Y H Kim, Liliana Goumnerova, Peter Black, Ching Lau, Jeffrey Allen, David Zagzag, James M Olson, Tom Curran, Cynthia Wetmore, Jaclyn A Biegel, Tomaso Poggio, Sayan Mukherjee, Ryan Rifkin, and Todd R Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415:436–42, 02 2002.