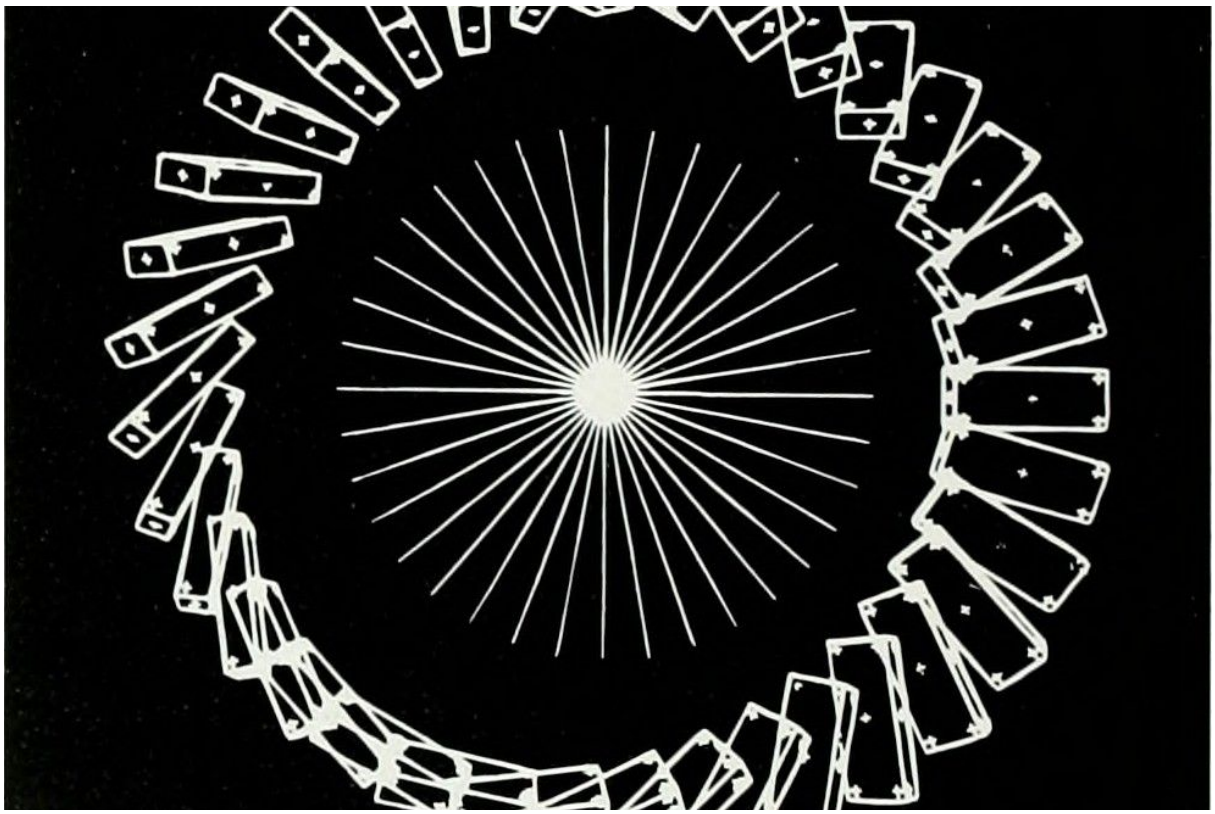


Projet : Manifold Learning

Étude comparative des méthodes de réduction de la dimension

Christopher Houbeiche, Sara Touzani, JB Miak

Objectifs

Dans un contexte d'effectuer une étude comparative des méthodes de réductions de dimension sur des manifolds nous avons été amené à travers ce projet à réaliser ce travail sur trois phases:

Simulation: de manifolds (S curv, Hémisphère, fish ball , ball , un cube et un twin peaks)

Estimation: où l'on a commencer par appliquer les méthodes qu'on connaît sur les manifolds , en ajustant leur paramètres pour avoir une meilleure visualisation de notre manifold en 2 dimensions

Comparaison: On comparait les modèles utilisés suivant un critère de variance (qu'on cherche à minimiser) des distances projetés Dy en 2 dimensions par rapport à la matrice de distance initiale Dx de notre manifold suivant la formule suivante:

$$E(Dx, Dy) = \sqrt{((\frac{1}{n}) \times \sum^{i=1..n} \sum^{j=1..n} (Dy_{i,j} - Dx_{moy})^2)}$$

avec: $Dx_{moy} = \frac{1}{n} \times \sum^{i=1..n} \sum^{j=1..n} Dx_{i,j}$

Matériels et méthodes

MDS:

Le MDS classique conserve les distances de manière globale. A utiliser lorsque les données sont sous forme de distance entre les observations.

ISOMAP:

Isomap s'inscrit dans la même démarche que MDS mais en remplaçant la distance euclidienne par la distance géodésique.

LLE:

Local Linear Embedding conserve les distances de manière locale. C'est une méthode qui s'est révélée efficace pour projeter les données en une dimension plus petite que la dimension intrinsèque.

PCA:

L'ACP va révéler les composantes qui ont une valeur propre supérieure à celle des autres. C'est une approche empirique car il faut repérer un coude dans les valeurs propres triées.

Sammon:

Méthodes de visualisation de données à fortes dimensions dans un espace réduit à 2 D

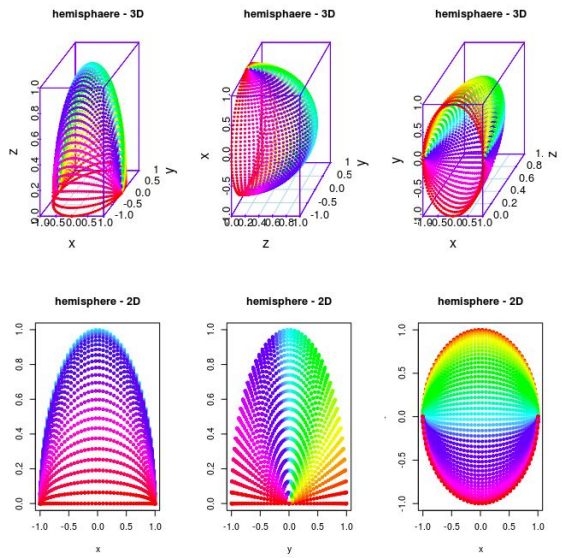
Kernel PCA:

L'ACP à noyau est une ACP non linéaire. Avec différent noyau "polynomiale, gaussein,..."

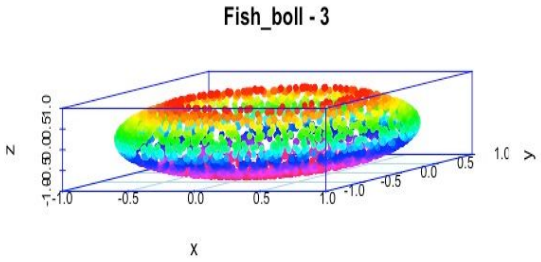
Données simulées

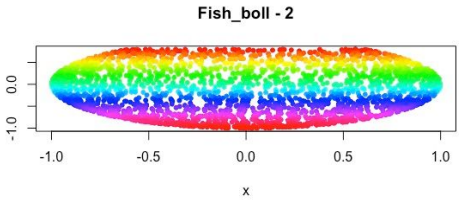
On a commencé le projet par des simulations de données qui seront par la suite utiliser dans la partie estimation et comparaison. On a simulé en total 3 jeu de données : Hémisphère, Fish Bowl et Twin Peaks.

Hémisphère:

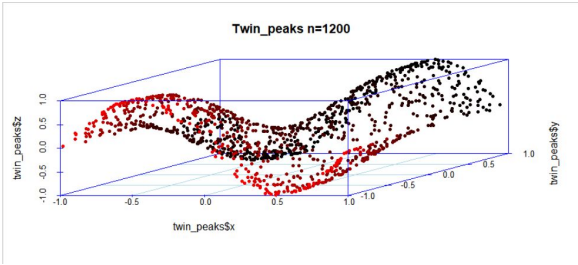


Fish Bowl:





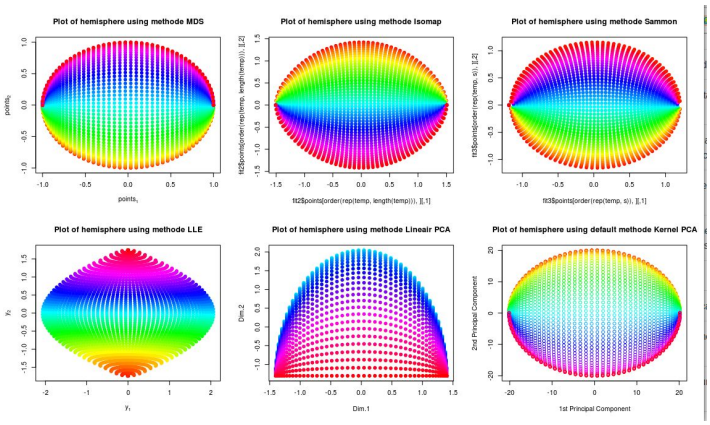
Twin Peaks:



Expériences numériques sur les données artificiels

Hémisphère

On a appliqué plusieurs méthodes sur notre manifold où il fallait faire des ajustement pour avoir une meilleur représentation où l'on a commencé par appliquer :

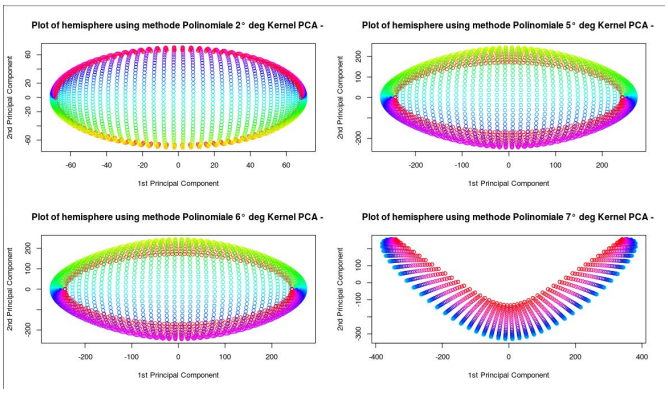


- MDS où les valeurs propres extraites montraient que la dimension estimé de cette dernière est bien ~ 2 ce qui se voit à travers sa représentation graphique qui préserve la distance et la variation au même temps.
- SOMAP où on a choisit de de préciser le voisinage epsilon=0.2 au lieu d'utiliser les k plus proche voisin ce qui nous a permis d'optimiser en temps d'exécution de la fonction mais en gardant la même qualité de représentation.

Pour la méthode Sammon lorsqu'on a injecter directement la matrice on a eu une erreur dû au fait que les données à partir de l'observation 2452 on la même valeur, du coup on a dû les enlever et rajouter un bruit avec jitter() avant d'appliquer la méthode Sammon

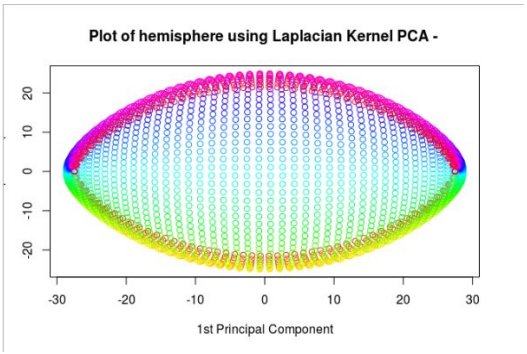
Comme Sammon LLE ne pouvait être appliqué qu'après avoir les points redondants

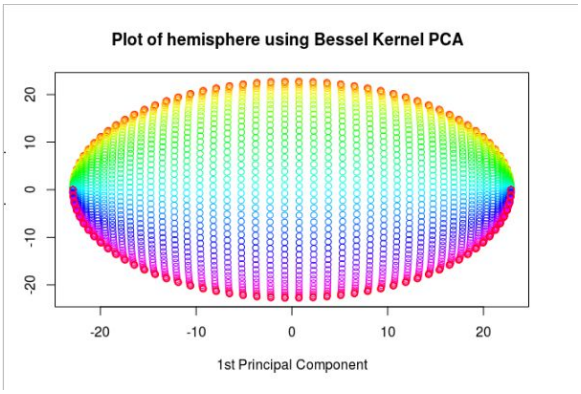
Pour le PCA linéaire on a pu observer que c'est pas la bonne méthode à appliquer sur ce type de données vu sa variation non linéaire



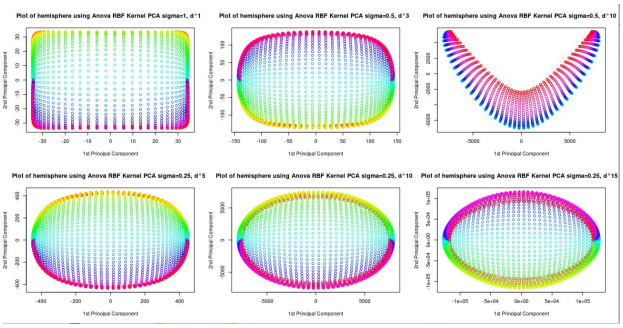
On a appliquer aussi la méthode du pca avec sa fonction kernel polynomiale où celle qui préserve la distance et la variation en second degré le plus est celle ayant deg=2, aussi lorsque on augmente le degré de la famille des polynômes on commence à avoir des points qui se chevauchent mais qui ne le sont pas dans la réalité et vers un moment on perds même la variabilité de notre manifold

la projection obtenue par le kpca utilisant le kernel laplacian et Bessel conserve le critère de la variation de l'hémisphère pour la première et la distance pour le deuxième mais il est surement pas le meilleurs parmi les autres méthodes pour cette forme



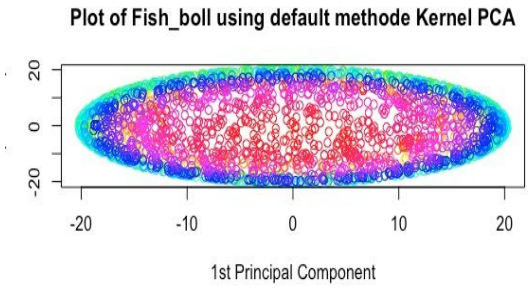


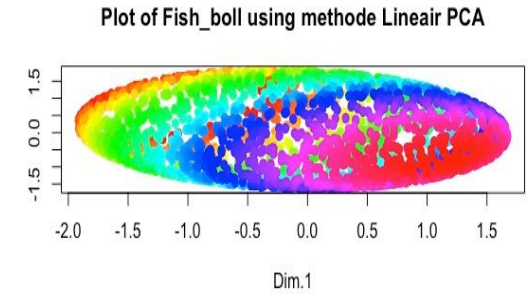
pour le kpca utilisant la fonction du kernel du Anova RBF : où il fallait diminuer sigma suffisamment pour garder la même forme initiale du manifold et en augmentant le deg

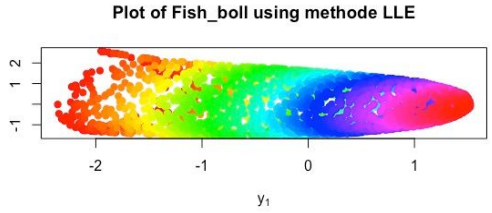


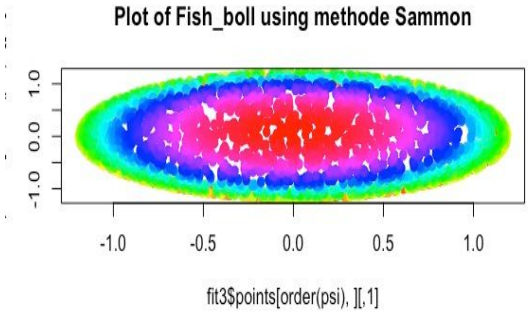
Fish Bowl

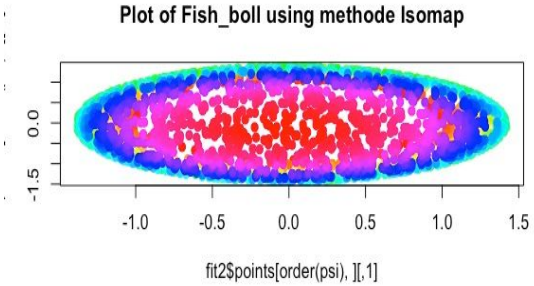
Ainsi on a suivi la même démarche que pour hemisphère et twinpeaks pour projeter fish bowl en 2D et on a obtenue :



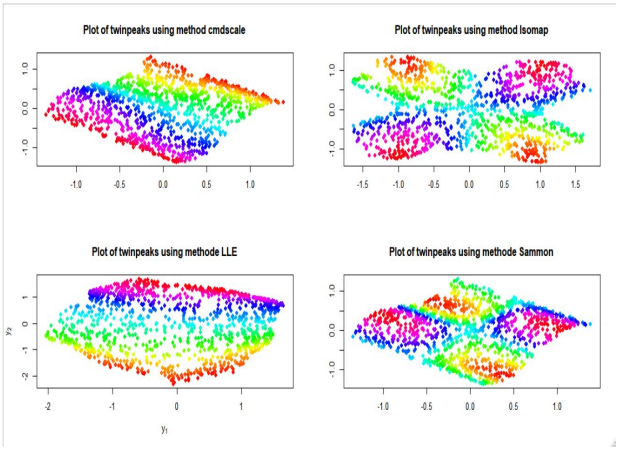








Twin Peaks



Les différentes méthodes appliqués de réduction de dimensionnalité pour les données twin peaks semble ne pas beaucoup déformées le nuage de points. Or, parmi les méthodes utilisées on peut en déduire visuellement que la meilleur méthode semble être celle de l'Isomap.

Isomap permet aussi de mettre les données sur un espace beaucoup plus réduit que les autres méthodes utilisées précédemment notamment les méthodes proche du PCA (MDS,Sammon) avec 2 valeurs propres significativement grands. Or, les axes principaux sont trois pour les méthodes de type PCA. On a utiliser aussi une méthode d'estimation du nombre optimale de plus proche voisin avec la fonction "calc_k" pour optimiser au plus le résultat du calcule de la fonction LLE.

Comparaison des méthodes

Appliquer des méthodes de réduction de dimension sur des manifold vient avec un inconvénient lequel de perdre de l'information , ainsi il faut préciser ce qui est important comme propriété à préserver entre:

- La connection entre les composants du manifold , Discontinuité - Trous (exemple: Projection de Mercator)
- Courbure de la forme
- Finesse de la variété de la forme
- Distance
- Angles , orientations …

Nous on a choisit de préserver plutôt la distance en premier degré , puis la variance , et la finesse de cette dernière.

On a choisit donc de calculer l'erreur moyenne par rapport à la distance initiale en se basant sur la formule citée sur "Objectifs"

Hemisphere

methode	e
SAMMON	1.185208e-01
MDS	1.406825e-01
ISOMAP	3.622506e-01
LLE	7.440596e-01
Lin-PCA	1.197044e+00
Def-K-PCA	2.608470e+01
Bessel-KPCA	2.959953e+01
Laplacian-KPCA	5.396685e+01
Spline-KPCA	6.527398e+01
Anova-RBF-sig1-d1-KPCA	6.560983e+01
Optimal Pol KPCA	1.063624e+02
Optimal-Anova-RBF-sig0.5-d3-KPCA	2.443255e+02
Optimal-Anova-RBF-sig0.5-d10-KPCA	7.144839e+02
Optimal-Anova-RBF-0.25-d10-KPCA	1.345819e+04
Optimal-Anova-RBF-0.25-d5-KPCA	1.499120e+04
Optimal-Anova-RBF-0.25-d15-KPCA	2.260950e+05

Ainsi pour notre forme hémisphère la meilleur méthode pour réduire la dimension de notre manifold en D=2 est la méthode de Sammon avec l'erreur moyenne la plus basse.

Fish Bowl

	LLE	Sammon_Non_Linear_Mapping	MDS	Isomap
1	0.7714736	0.3468453	0.3538747	0.4363001

On a appliqué différent type de modèle sur les données Fish Bowl. La méthode qui semble déformer au plus le nuage de point est celle de la LLE qui change significativement la forme initiale du nuage de point.

Twin Peaks

	CMDScale	Sammon_Non_Linear_Mapping	MDS	Isomap
1	0.4250825	0.4189539	0.4250825	0.3700276

Après une évaluation des différentes erreurs RMSE sur les distances on peut voir que notre hypothèse est bien verifier. Isomap est meilleur modèle avec l'erreur la plus basse.

Réduction de dimension de Données Réelle :

en se basant sur la décomposition des valeurs singulière à travers la formule :

$$\tilde{x} = \mathbf{z}^{1:k} V^{1:k^T}$$

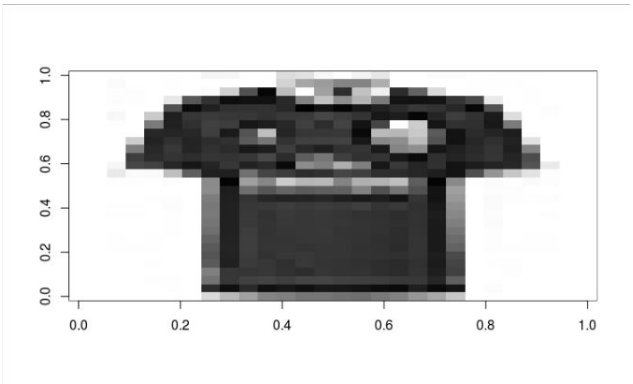
où : k est notre fenêtre de réduction(le nombre de dimension à garder)

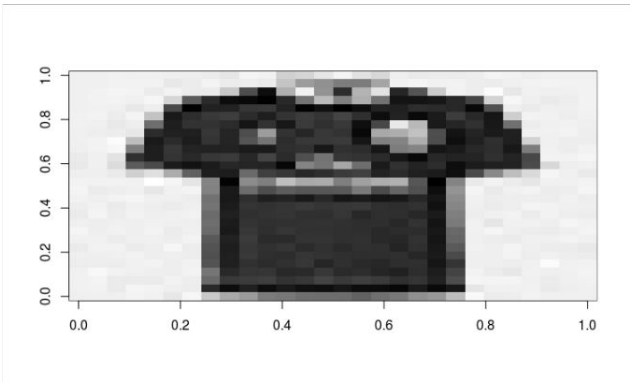
z : sont les coordonnées des composantes principales dans l'espace pivoté

v: la matrice de vecteurs propres

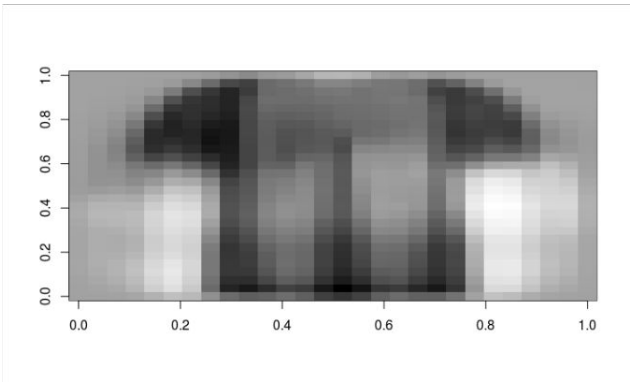
ainsi cette réduction est réversible en ajoutant la matrice moyenne de toutes les donnés étudiés .

Ainsi pour l'image prise de la dataset fashion mnist :

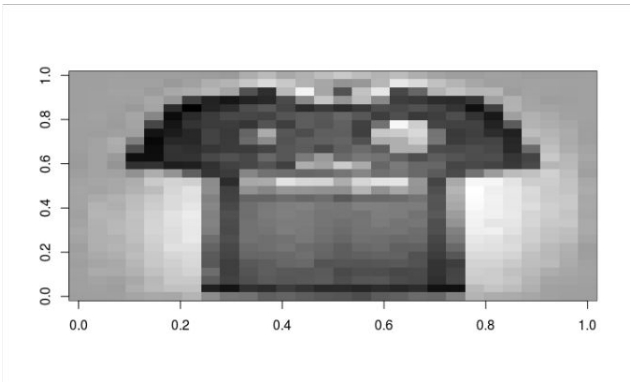




on a obtenu pour k=10 dimention préservé :



pour k=50



ainsi l'image retournée après réduction pour k=50 est celle qui suit: