Réduction et transformation de données

Daniel Aloise <daniel.aloise@polymtl.ca>

14 septembre 2021

Réduction de données

- Lorsque la taille des données est plus petite, il est plus facile et performant d'appliquer des algorithmes sophistiqués (et complexes).
- Plus facile d'interpréter les résultats et de les visualiser.
- La réduction des données peut se faire en termes du nombre de lignes (enregistrements) et/ou en termes du nombre de colonnes (dimensions).
- La réduction de données entraîne toujours une perte d'information.

Réduction de données

- Échantillonnage de données : les enregistrements sont échantillonnés pour créer une base de données beaucoup plus petite.
- Sélection d'attributs : seul un sous-ensemble des attributs est utilisé dans le processus analytique.
- Réduction de dimension : les corrélations entre les lignes ou les colonnes sont exploitées pour représenter les données dans une dimension plus petite.

Échantillonnage

- Une fraction de données est sélectionnée et retenue pour l'analyse.
- Dans l'échantillonnage biaisé, certains enregistrements sont plus probablement retenus en raison de leur plus grande importance pour l'analyse.
  - ex. certains types de données temporelles
- Dans l'échantillonnage stratifié, les données sont d'abord divisées en ensemble de strates souhaitées, puis l'échantillonnage est fait indépendamment à partir de chacune des strates en fonction de proportions prédéfinies
  - ex. sondage sur le style de vie des individus d'une population ; un échantillon aléatoire de 1 million de participants peut ne pas capturer un milliardaire.

Sélection d'attributs

Sélectionner les colonnes

- *Learning expert system for medical diagnosis* (LEXMED) contient les données de $n = 473$ patients.
- Disponible ici.

Var. num.	Description	Values
1	age	continuous
2	sex (1=male, 2=female)	1,2
3	pain quadrant 1	0,1
4	pain quadrant 2	0,1
5	pain quadrant 3	0,1
6	pain quadrant 4	0,1
7	local muscular guarding	0,1
8	generalized muscular guarding	0,1
9	rebound tenderness	0,1
10	pain on tapping	0,1
11	pain during rectal examination	0,1
12	axial temperature	continuous
13	rectal temperature	continuous
14	leucocytes	continuous
15	diabetes mellitus	0,1
16	appendicitis	0,1

Source : Ertel, 2011

Sélection d'attributs

$X_1 = (26, 1, 0, 0, 1, 0, 1, 0, 1, 1, 0, 37.9, 38.8, 23100, 0, 1)$
$X_2 = (17, 2, 0, 0, 1, 0, 1, 0, 1, 1, 0, 36.9, 37.4, 8100, 0, 0)$

- moyenne :

$$\overline{X^s} = \frac{1}{n}\sum_{i=1}^{n} X_i^s$$

- écarte type :
  variance

$$\sigma(X^s) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i^s - \overline{X^s})^2}$$

- covariance

mesure le synchronisme des attributs ou individus, c-à-d si deux attributs augmentent en mm temps
attributs synchronisés en mm temps en haut de la moyenne ou en bas de la moyenne

$$Cov(X^s, X^t) = \frac{1}{n}\sum_{i=1}^{n}(X_i^s - \overline{X^s}) \cdot (X_i^t - \overline{X^t})$$

Matrice de covariance

- En calculant la covariance entre chaque pair d'enregistrements (ou chaque pair d'attributs), nous obtenons la matrice de covariance qui contient chaque paire d'attributs
- Possible transformation :
  - centralisation par la moyenne :

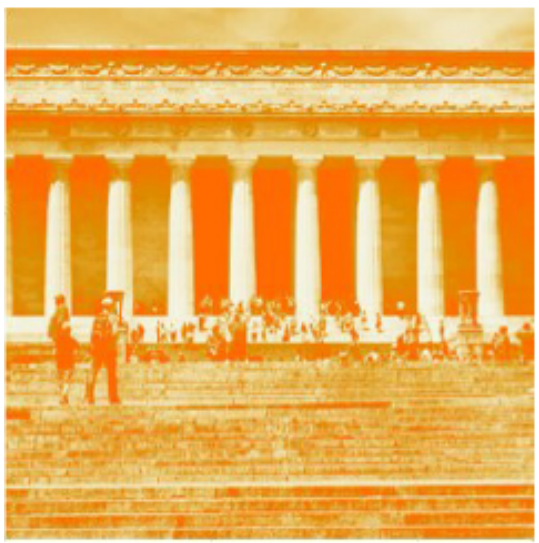
$$X_i^s \leftarrow X_i^s - \overline{X}^s \qquad \forall i = 1, \ldots, n; \ \forall s = 1, \ldots, d.$$

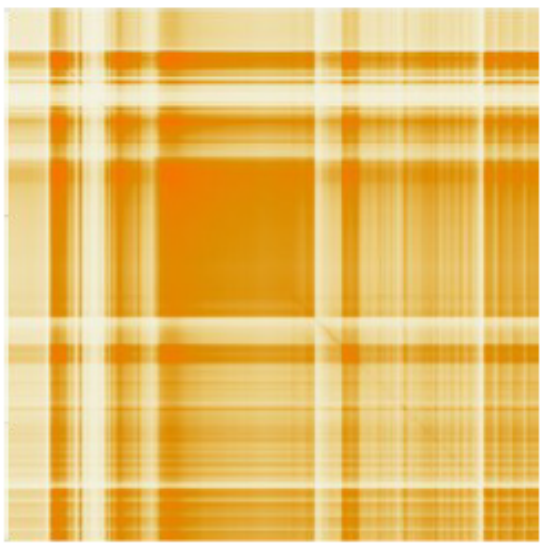
- $\Sigma = \frac{X^\top \cdot X}{n}$ est une matrice $d \times d$ de produits scalaires, mesurant "le synchronisme" entre les attributs.
- $\Sigma = \frac{X \cdot X^\top}{d}$ est une matrice $n \times n$ de produits scalaires, mesurant "le synchronisme" entre les enregistrements.

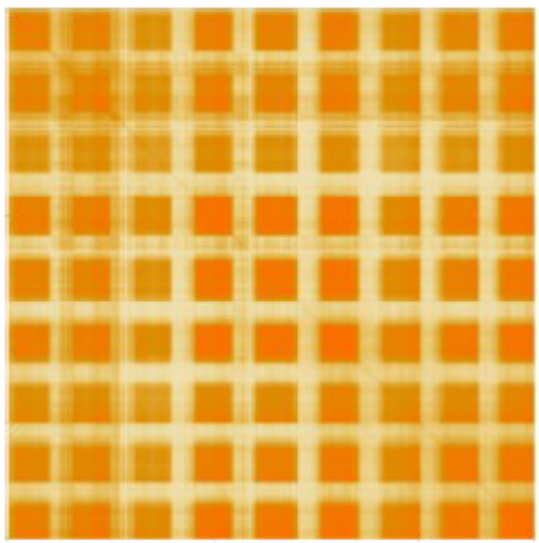
enregistrements: ce sont les individus ou les lignes

Matrice de covariance

Quelle image correspond à $X \cdot X^\top$? Et à $X^\top \cdot X$?







Source : Skiena, 2017

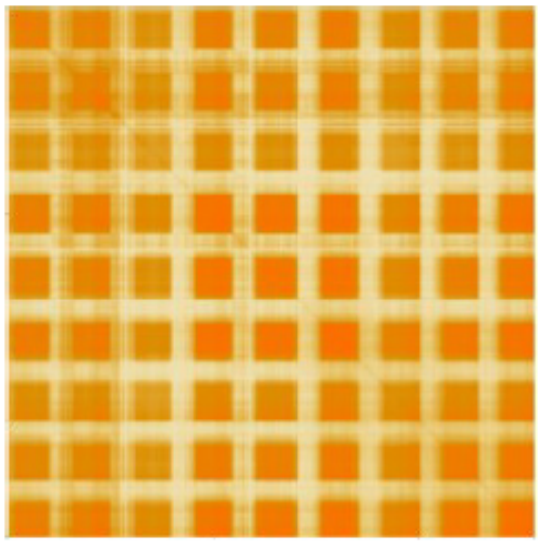
Plus c'est sombre, plus la covariance est élevée

Matrice de covariance

Quelle image correspond à $X \cdot X^\top$? Et à $X^\top \cdot X$?

Les lignes

Source : Skiena, 2017



Les colonnes

$X \cdot X^\top$

$X^\top \cdot X$

Sélection des attributs

- Coefficient de corrélation : tenir compte de l'écart type de nos attributs

$$Corr(X^s, X^t) = \frac{Cov(X^s, X^t)}{\sigma(X^s) \cdot \sigma(X^t)}$$

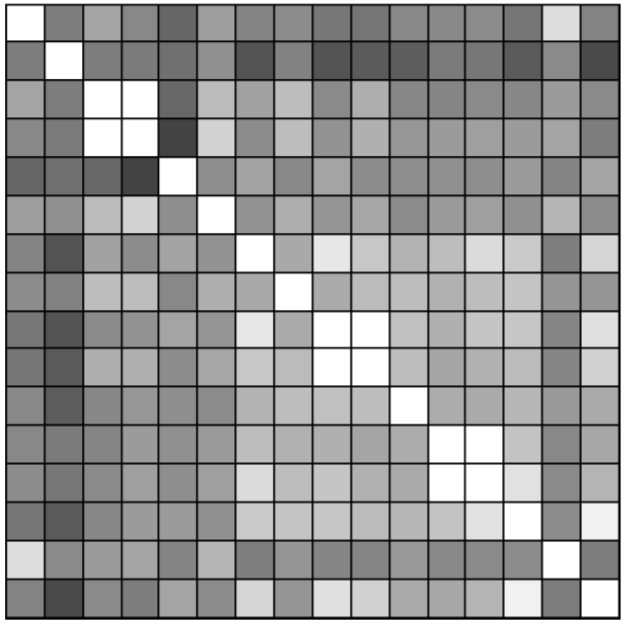
on peut inverser s et t sans problème

- Matrice de corrélation pour les 16 attributs d'appendicite mesurés dans 473 cas. symétrique

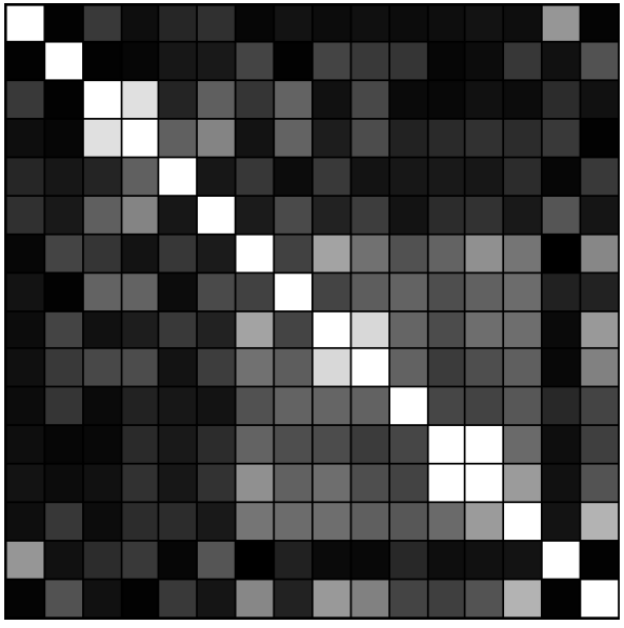
1.	−0.009	0.14	0.037	−0.096	0.12	0.018	0.051	−0.034	−0.041	0.034	0.037	0.05	−0.037	0.37	0.012
−0.009	1.	−0.0074	−0.019	−0.06	0.063	−0.17	0.0084	0.17	−0.14	−0.13	−0.017	−0.034	0.14	0.045	−0.2
0.14	−0.0074	1.	0.55	−0.091	0.24	0.13	0.24	0.045	0.18	0.028	0.02	0.045	0.03	0.11	0.045
0.037	−0.019	0.55	1.	−0.24	0.33	0.051	0.25	0.074	0.19	0.087	0.11	0.12	0.11	0.14	−0.0091
−0.096	−0.06	−0.091	−0.24	1.	0.059	0.14	0.034	0.14	0.049	0.057	0.064	0.058	0.11	0.017	0.14
0.12	0.063	0.24	0.33	0.059	1.	0.071	0.19	0.086	0.15	0.048	0.11	0.12	0.063	0.21	0.053
0.018	−0.17	0.13	0.051	0.14	0.071	1.	0.16	0.4	0.28	0.2	0.24	0.36	0.29	−0.0001	0.33
0.051	0.0084	0.24	0.25	0.034	0.19	0.16	1.	0.17	0.23	0.24	0.19	0.24	0.27	0.083	0.084
−0.034	−0.17	0.045	0.074	0.14	0.086	0.4	0.17	1.	0.53	0.25	0.19	0.27	0.27	0.026	0.38
−0.041	−0.14	0.18	0.19	0.049	0.15	0.28	0.23	0.53	1.	0.24	0.15	0.19	0.23	0.02	0.32
0.034	−0.13	0.028	0.087	0.057	0.048	0.2	0.24	0.25	0.24	1.	0.17	0.17	0.22	0.098	0.17
0.037	−0.017	0.02	0.11	0.064	0.11	0.24	0.19	0.19	0.15	0.17	1.	0.72	0.26	0.035	0.15
0.05	−0.034	0.045	0.12	0.058	0.12	0.36	0.24	0.27	0.19	0.17	0.72	1.	0.38	0.044	0.21
−0.037	−0.14	0.03	0.11	0.11	0.063	0.29	0.27	0.27	0.23	0.22	0.26	0.38	1.	0.051	0.44
0.37	0.045	0.11	0.14	0.017	0.21	−0.0001	0.083	0.026	0.02	0.098	0.035	0.044	0.051	1.	−0.0055
0.012	−0.2	0.045	−0.0091	0.14	0.053	0.33	0.084	0.38	0.32	0.17	0.15	0.21	0.44	−0.0055	1.

On peut slm se focuser sur ce qu'il y a en dessous de la diagonale

Source : Ertel, 2011



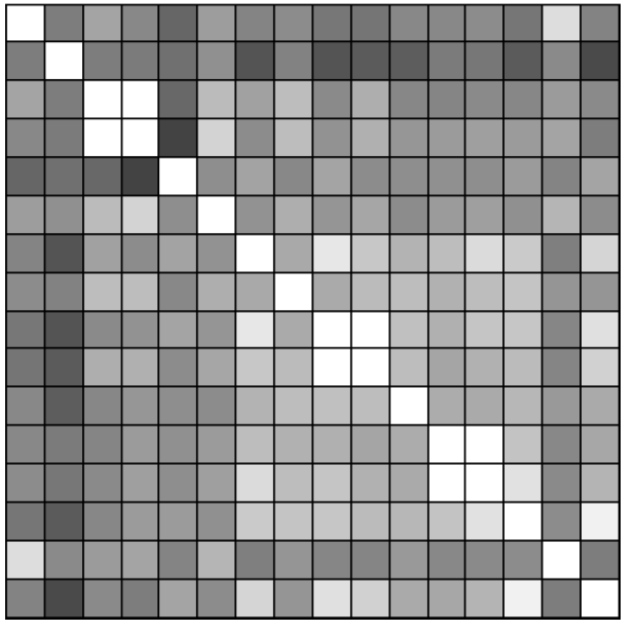
$Corr(X^s, X^t) = -1$ (noir)
$Corr(X^s, X^t) = 1$ (blanc)



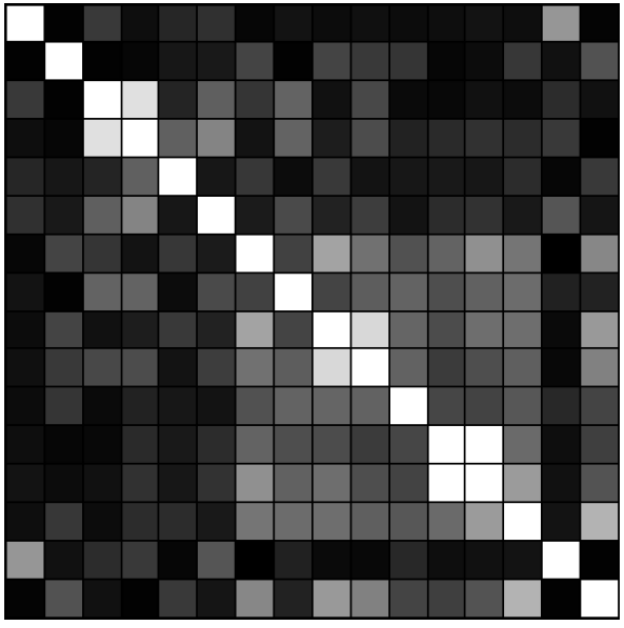
$|Corr(X^s, X^t)| = 0$ (noir)
$|Corr(X^s, X^t)| = 1$ (blanc)

- Les attributs 7, 9, 10 et 14 sont les plus importants pour diagnostiquer une appendicite - attribut 16 (pourquoi ?)

  On regarde la dernière colonne et on voit que c'est les plus pâles



$Corr(X^s, X^t) = -1$ (noir)
$Corr(X^s, X^t) = 1$ (blanc)



$|Corr(X^s, X^t)| = 0$ (noir)
$|Corr(X^s, X^t)| = 1$ (blanc)

- Les attributs 7, 9, 10 et 14 sont les plus importants pour diagnostiquer une appendicite - attribut 16 (pourquoi ?)
- En plus, on pourrait garder soit l'attribut 9 ou l'attribut 10 (pourquoi ?)

Ils sont corrélés (pâles), donc redondants

Sélection d'attributs

- La matrice de corrélation permet d'évaluer seulement des pairs d'attributs.
- On pourrait évaluer des triples ou des corrélations pour des sous-ensembles d'attributs plus grands.
- Cependant, le nombre de sous-ensembles est exponentiel.
- Par conséquent, en pratique, la plupart des méthodes de sélection évaluent les attributs indépendamment les uns des autres.
- Les méthodes sont différentes en fonction des données disponibles (avec ou sans étiquettes).

Filtrage

- Classe chaque attribut en fonction d'une mesure univariée.
- Sélectionne les attributs avec les mesures les plus élevées.
- La mesure doit refléter le pouvoir discriminant de chaque attribut.
- Ex. score de Fisher :

$$F(s) = \frac{\sum_{j=1}^{k} p_j(\overline{X_j^s} - \overline{X^s})^2}{\sum_{j=1}^{k} p_j(\sigma(X_j^s))^2}$$

où p_j est la fraction de données appartenant à la classe j

Filtrage corrélé

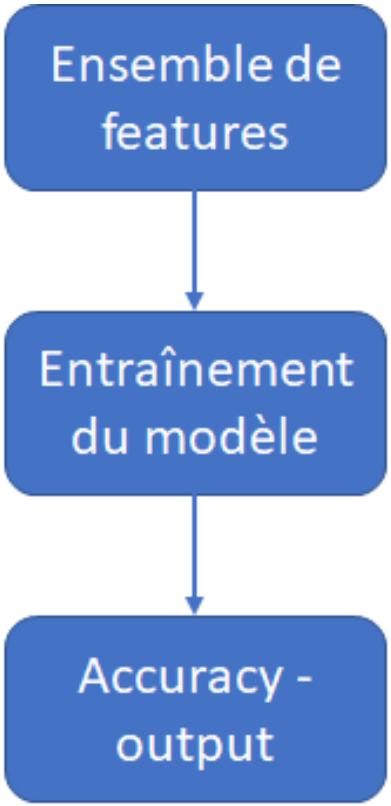
Algorithme itératif

- sélectionnez un attribut s (selon une métrique quelconque).
- vérifiez la corrélation de s avec les attributs déjà sélectionnés (par pair d'attributs).
- Si la somme de ces corrélations dépasse un seuil, enlève s de la sélection et STOP !

Filtrage

- Avantages :
  - Rapide à calculer.
- Désavantages :
  - Un attribut qui n'est pas "utile" tout seul peut être très utile lorsqu'il est combiné avec d'autres.
- Discriminant linéaire de Fisher : filtrage des combinaisons linéaires des attributs.

Emballage



- Influencé par le modèle choisi (biaisé).
- Coûteux en temps de calcul.
- Si nous n'avons pas d'étiquettes, on exécute une méthode de *clustering* pour en avoir.

Emballage

Sequential Forward Selection

$S \leftarrow \emptyset$;
while critères d'arrêt non satisfaits **do**
  for attribut $j = 1$ to d **do**
    $S' \leftarrow S \cup \{j\}$
    Entraîne le modèle M avec S'
    Calcule la performance du modèle
  end for
  $S \leftarrow S^*$, où S^* est l'ensemble des attributs avec la plus grande performance de classification
end while

Réduction de dimension

Décomposition par valeur propres

Toute matrice A carrée symétrique de taille $n \times n$ peut être décomposée de la façon suivante :

$$A = \sum_{i=1}^{n} \lambda_i U_i U_i^T$$

où U_i est le vecteur propre associé à la valeur propre λ_i

Remarque que les produits plus importants sont ceux associés aux valeurs propres les plus grandes.

Analyse de composantes principales

- La matrice de covariance $\Sigma = \frac{X^\top X}{n}$ est symétrique $(d \times d)$ et semidéfinie positive.
- Elle peut donc être décomposée par :

$$\Sigma = \sum_{i=1}^{d} \lambda_i U_i U_i^\top$$

où tous les $\lambda_i \geq 0, i = 1, \ldots, d$.

- Les vecteurs U_i, $i = 1, \ldots, d$ sont nommés composantes principales pour la PCA.
- Ils sont triés en ordre décroissant par rapport à λ_i.

Analyse de composantes principales

- On dénote X' la matrice $n \times d$ obtenue par :

$$X' = X \cdot U$$

- Pour la PCA, seulement $k \ll d$ colonnes de X' présentent des valeurs qui varient de façon significative.

- On peut prouver que les vecteurs propres U_i représentent de solutions orthogonales successives au problème de la maximisation de la variance $v^T \Sigma v$ le long d'une direction unitaire v.

Réduction de dimension

- Le problème de réduction de dimension d'une matrice X de données peut aussi être vu comme celui de factoriser une matrice :

$$X = B \cdot C$$

où B et C sont de taille $n \times k$ et $k \times d$

- Si $k < \min(n, d)$, B et C réduisent X.

Décomposition par valeurs singulières

- La SVD d'une matrice X de taille $n \times d$ la factorise comme :

$$X = WDV^T$$

où D est une matrice diagonale rectangulaire $n \times d$, W est de taille $n \times n$ et V est de taille $d \times d$.

- Les matrices W et V sont orthogonales.
- Donc, WD pondère chaque colonne de W par D, de même pour DV^T.
- Le fait de ne conserver que les lignes/colonnes avec des poids d_{ii} importants (singular values) nous permet de compresser X avec relativement peu de perte.

Décomposition par valeurs singulières

- Soit A et B deux vecteurs de taille $n \times 1$ et $1 \times m$, resp.
- Le produit externe des vecteurs donne une matrice :

$$P = A \otimes B \qquad P[i,j] = A[i]B[j]$$

- Notre matrice X peut être exprimée par la somme des produits externes de la SVD avec les termes : $(WD)_k$ et $(V^T)_k$.

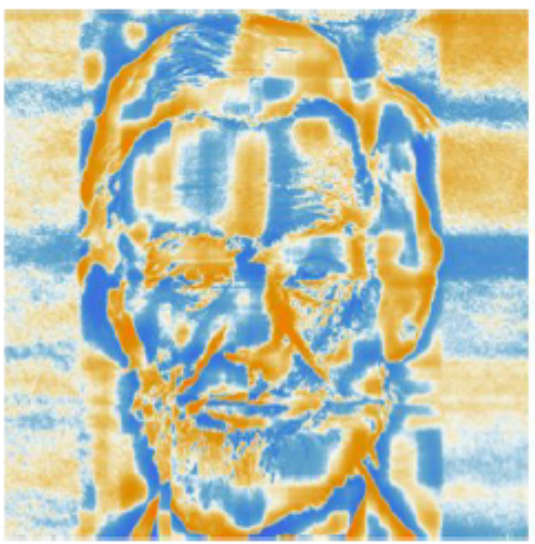
$$X = WDV^T = \sum_k WD_k \otimes V_k^T$$

- En additionnant seulement les plus grands produits matriciels, on obtient une approximation de X.

Exemple

- Image compressée avec 5 et 50 valeurs singulières (d'un total de 512).





Source : Skiena, 2017

Multidimensional scaling (MDS)

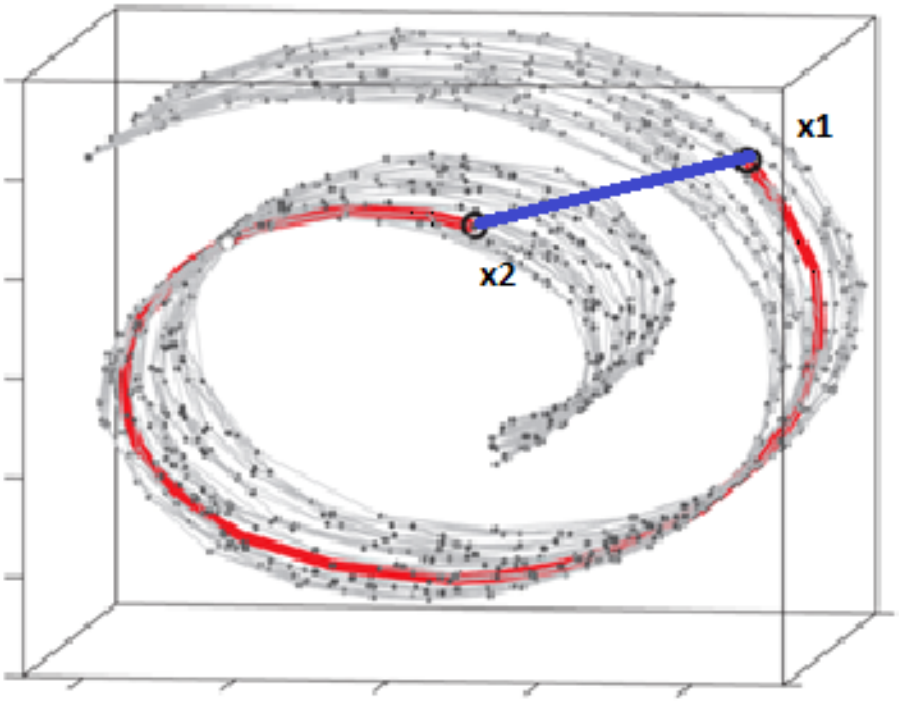
- MDS projette un espace de grande dimension d'origine sur un espace de petite dimension en préservant les distances entre les paires.
- *input* : matrice de distances D de taille $n \times n$ obtenue dans l'espace original, et la dimension $k < d$ de la projection.
- *output* : une configuration $Y_1, \ldots, Y_n$ de n points dans $\mathbb{R}^k$.
- *metric* : MDS minimise

$$\min_{Y \in \mathbb{R}^k} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} (D_{ij} - \|Y_i - Y_j\|)^2$$

- Un algorithme de descente du gradient peut être utilisé pour optimiser MDS.

ISOMAP

- Plusieurs méthodes ont récemment été proposées pour la réduction non linéaire de dimensions.
- L'idée est que les données font partie d'un *manifold* non linéaire de dimension plus faible dedans un espace de plus grande dimension.



on calcule la distance bleue alors qu'on veut la rouge

ISOMAP

- L'algorithme ISOMAP procède en trois étapes :
  1. Pour chaque $X_i, i = 1, \ldots, n$, nous trouvons ses voisins situés à l'intérieur d'une petite distance euclidienne de X_i.
  2. Nous construisons un graphe avec une arête entre tous les points voisins.
  3. La distance géodésique entre deux points quelconques est ensuite approchée par le chemin le plus court entre les points dans le graphe.
  4. Enfin, MDS est appliqué aux distances obtenues pour produire une représentation en dimension faible.

t-distributed stochastic neighbor embedding (t-SNE)

- Chaque donnée multidimensionnelle est modélisée par un autre à plus petite dimension.
- t-SNE préserve la structure locale en maintenant autant que possible la structure globale intacte.
- Modèle probabiliste : Les enregistrements similaires sont modélisés par des points voisins et des données dissemblables sont modélisées par des points éloignés avec une probabilité élevée.
- Cela dans l'espace original et dans l'espace réduit.

t-distributed stochastic neighbor embedding (t-SNE)

- t-SNE marche de la façon suivante :
  - Une distribution de probabilité sur des paires de données est construite de telle manière que les données similaires ont une forte probabilité d'être sélectionnées ensemble.
  - On cherche une distribution semblable pour les données mappées en dimension réduite.
  - Pour y achever, la divergence de Kullback-Leibler est minimisée entre les deux distributions (problème d'optimisation non convexe).

Transformation spectrale

- Graphe de similarité $\Rightarrow$ données multidimensionnelles.
- Conserve la structure de similarité d'un point de vue local.
- Très utile étant donné l'importance des graphes comme structure de représentation.

Transformation spectrale

- Soit $G(N, A)$ un graphe non dirigé.
- On assume que $|N| = n$.
- Une matrice symétrique W de taille $n \times n$ contient les similarités entre les noeuds liés par une arête.
- Cette matrice est creuse.
- On veut incorporer les noeuds de ce graphe dans un espace de dimension d afin que la structure de similarité des données soit préservée.

Transformation spectrale

- Tout d'abord, discutons du cas plus simple de la mise en correspondance des n noeuds sur un ensemble de valeurs $y_1, \ldots, y_n$ à une dimension.
- Il n'est pas souhaitable que des noeuds connectés avec des arêtes de poids élevé (très similaires) soient mappés sur des points éloignés de cette ligne.
- On veut minimiser alors :

$$O = \min \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} (y_i - y_j)^2$$

Transformation spectrale

- O peut être réécrit en termes de la matrice Laplacienne L de W.

$$O = 2y^T Ly^T$$

- où D est une matrice diagonale telle que $D_{ii} = \sum_{j=1}^{n} w_{ij}$ et $L = D - W$.

- On peut démontrer que O est minimisé avec $y = (y_1 \dots y_n)^T$ égal au plus petit vecteur propre de la relation $D^{-1}Ly = \lambda y$.

- Pourtant, le plus petit vecteur propre associé n'est pas informatif. ($\lambda = 0$)

- On prend donc le vecteur propre associé à la deuxième plus petite valeur propre.

Transformation spectrale

- La généralisation au cas k-dimensionnel est relativement simple.

- Les autres vecteurs propres $U_2, U_3, \ldots, U_{k+1}$ associés aux valeurs propres $\lambda_2 \leq \lambda_3 \leq \ldots \leq \lambda_{k+1}$ forment les coordonnées de la matrice multidimensionnelle de taille $n \times k$.