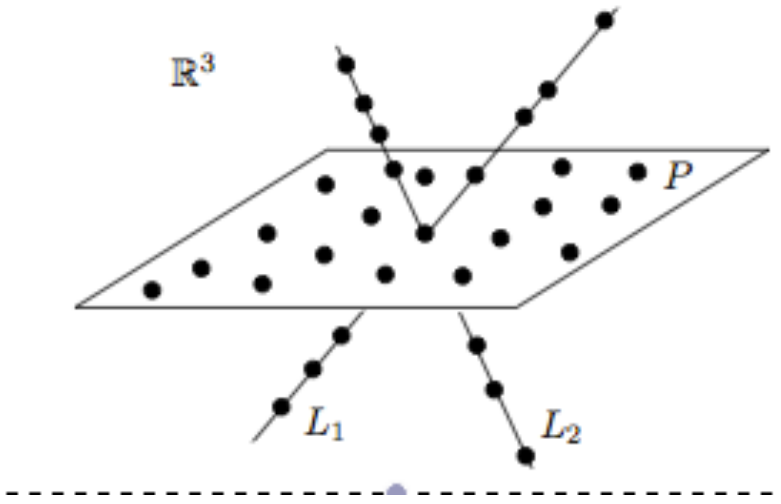
Lecture de l'article

Robust Subspace Clustering

Catherine d'Aubigny

Objectif



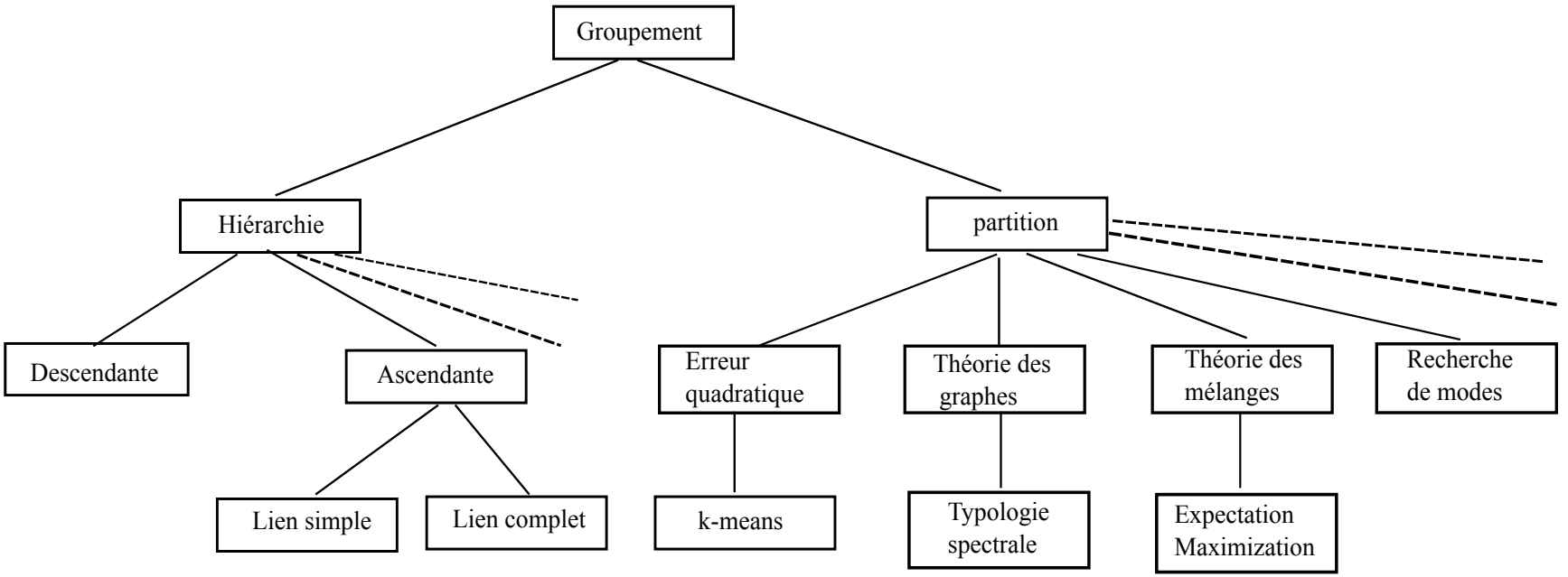
N points $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}\}$ dans une union de K sous-espaces de $\mathbb{R}^p$

- identifier les sous-espaces
- donner une base
- segmenter les points en les affectant aux sous-espaces

Robust Subspace Clustering

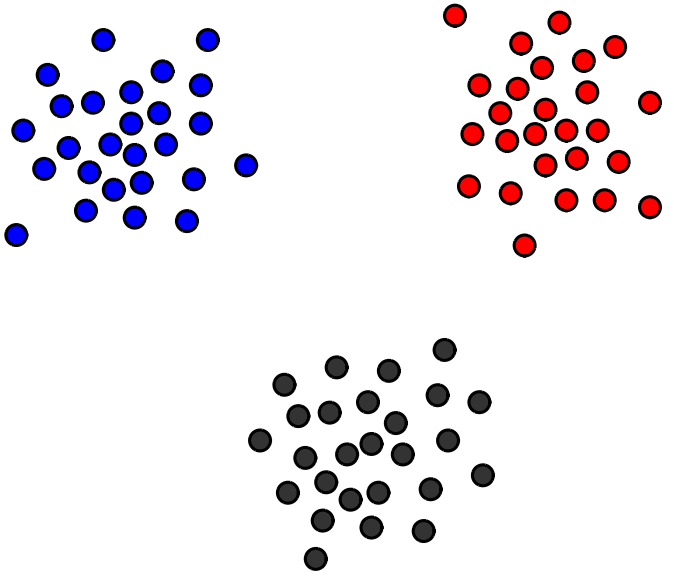
trois mots :

- Clustering $\Rightarrow$ méthodes de classification
- Subspace $\Rightarrow$ méthodes de représentation des données
- Robust $\Rightarrow$ données bruitées et données abérrantes

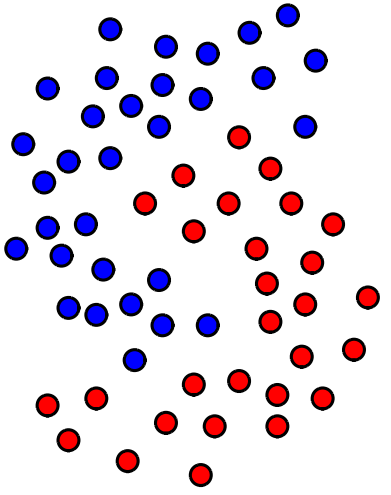

Groupement
Hiérarchie
partition
Descendante
Ascendante
Erreur quadratique
Théorie des graphes
Théorie des mélanges
Recherche de modes
Lien simple
Lien complet
k-means
Typologie spectrale
Expectation Maximization

Vue rapide

- basée sur une mesure de dissimilarité entre objets et sur l'heuristique : les classes sont constituées d'objets qui se ressemblent
  - méthodes de classification hiérarchique
  - méthodes de partitionnement : K-means, spectral clustering, ...
- basée sur un modèle décrivant le processus qui génère les données : modèle de mélange
- ....



Compacité



Connexité

K-means

méthode itérative qui à partir d'un choix initial de centres, alterne la construction des classes en agrégeant les objets autour des centres et en recalculant de nouveaux centres comme centres de gravité des classes

- dissimilarité = distance euclidienne ordinaire
- hétérogénéité d'une classe C : $\sum_{i \in C} d^2(i, G_C)$
- optimalité : à chaque étape, l'hétérogénéité des classes diminue
- construit des classes sphériques

Modèle de mélange

données complétées : $(\mathbf{x}_i, \mathbf{z}_i)$, avec $\mathbf{x}_i \in \mathbb{R}^p$, et $\mathbf{z}_i \in \{0,1\}^K$
modèle statistique :

- $\mathbf{z}_i$ loi multinomiale de paramètre $\pi_1, \pi_2, \ldots, \pi_K$ paramètres du mélange
- $f(\mathbf{x}_i) = \sum_k \pi_k f_k(\mathbf{x}_i, \boldsymbol{\theta}_k)$

solution :

1. Estimation de $\boldsymbol{\theta}$ à l'aide de l'algorithme de maximisation *EM*
2. Estimation de $\mathbf{Z}$ à l'aide de la procédure *MAP* (Maximum A Posteriori)

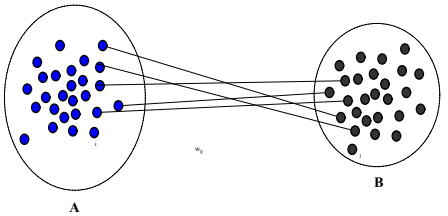
K-means et modèle de mélange sont équivalentes lorsque $f_k(\mathbf{x}_i, \boldsymbol{\theta}_k)$ est la densité de la loi $\mathcal{N}(\boldsymbol{\mu}_k, \sigma^2 \mathbf{I}_p)$

spectral clustering

le principe général

- Les N objets à classer sont représentés comme les sommets d'un graphe complet $G = (V, E)$
  classifier $\rightarrow$ partitionner le graphe en composantes connexes
- la similarité entre objets $\rightarrow$ poids $w_{ij} \in \mathbb{R}^{N \times N}$ des arêtes
  matrice de similarité du graphe = $\mathbf{W}$
  $\mathbf{W}$ est une matrice symétrique à termes non négatifs

coupure



coupure :

$$Cut(A,B) = \sum_{i \in A, j \in B} w_{ij}$$

$$Vol(A) = \sum_{i \in A} \sum_{j=1}^{N} w_{ij} = \sum_{i \in A} w_{i+}$$

coupure normalisée :

$$Ncut(A,B) = Cut(A,B)\left(\frac{1}{Vol(A)} + \frac{1}{Vol(B)}\right)$$

Méthode de la coupure maximale

On cherche 2 classes A et B dont la coupure est minimum $\Rightarrow$ la fonction objectf :

$$J_{MinCut} = Cut(A, B) \qquad \text{ou plutôt} \qquad J_{MinNcut} = Ncut(A, B)$$

problème NP-complet !
Le spectral clustering est une relaxation de ce problème

spectral clustering

$$\mathbf{q} = (q_1, q_2, \ldots, q_N)$$

ou

$$q_i = \begin{cases} \frac{1}{Vol(A)} & \text{si } i \in A, \\ -\frac{1}{Vol(B)} & \text{si } i \in B. \end{cases}$$

$$\mathbf{q}^t \mathbf{L} \mathbf{q} = \sum_{i \in A} \sum_{j=1}^{N} w_{ij} \left(\frac{1}{Vol(A)} + \frac{1}{Vol(B)} \right)^2$$

où $\mathbf{L} = \mathbf{D}_w - \mathbf{W}$, $\mathbf{L}$ Laplacian combinatoire et $\mathbf{D}_w = diag(w_{i+})$

$$\mathbf{q}^t \mathbf{D}_w \mathbf{q} = \left(\frac{1}{Vol(A)} + \frac{1}{Vol(B)} \right)^2$$

spectral clustering

$$J_{MinNcut} = \frac{\mathbf{q}^t \mathbf{L} \mathbf{q}}{\mathbf{q}^t \mathbf{D}_w \mathbf{q}}$$

la solution est obtenue par le calcul d'éléments propres :

- laplacien combinatoire : $\mathbf{L}\mathbf{q} = \lambda \mathbf{D}_w \mathbf{q}$
- laplacien normalisé : $\mathbf{D}_w^{-1/2} \mathbf{L} \mathbf{D}_w^{1/2} \mathbf{q} = \lambda \mathbf{q}$

Résultat : le nombre de valeurs propres nulles est égal au nombre de composantes connexes du graphe

Construction des classes

1. calcul des éléments propres
2. représentation des points à l'aide des K vecteurs propres associés au K plus petites valeurs propres
3. classification des points dans cet espace à l'aide d'un K-means

nécessité d'une étape de classification
choix de K ?

robuste

données bruitées
données : N points $\{\mathbf{y}_1, \ldots, \mathbf{y}_N\}$ de $\mathbb{R}^p$ avec

$$\mathbf{y}_i = \mathbf{x}_i + \mathbf{z}_i$$

où

$$\mathbf{x}_i \in \cup_k S_k \text{ et } \mathbf{z}_i \text{ iid de loi } \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}_p)$$

et S_k sous-espace de $\mathbb{R}^p$ de dimension d_k

les $\mathbf{x}_i$ sont non observés et les $\mathbf{z}_i$ modélisent le bruit

schéma général du subspace clustering

trois étapes :

1. définir une mesure de similarité entre les points
2. construire une typologie à l'aide de la méthode de spectral clustering
3. estimer les sous-espaces représentant chaque classe

quelle dissimilarité ?

cas des données non bruitées
idée : exprimer chaque vecteur $\mathbf{x}_i$ comme une combinaison parcimonieuse des autres vecteurs. On résoud le problème de minimisation :

$$\min \|\beta_i\|_1 \text{ sous les contraintes } \mathbf{x}_i = \mathbf{X}\beta_i \text{ et } \beta_i^i = 0$$

Heuristique fondée si on peut supposer que la représentation la plus parcimonieuse de $\mathbf{x}_i$ sélectionnera uniquement les vecteurs qui appartiennent au même sous-espace que $\mathbf{x}_i$.
On pose

$$\mathbf{B} = [\beta_1, \ldots, \beta_N]$$

Alors matrice de similarité : $\mathbf{W} = |\mathbf{B}| + |\mathbf{B}|^t$

données bruitées

$$\mathbf{x}_i = \mathbf{X}\beta_i \Rightarrow \mathbf{y}_i = \mathbf{Y}\beta_i + \mathbf{u}_i \quad (\mathbf{u}_i = \mathbf{z}_i - \mathbf{Z}\beta_i)$$

régression parcimonieuse $\rightarrow \hat{\beta}_i \rightarrow \mathbf{W}$
même étape 2
ajout d'une procédure de débruitage lors l'étape 3

algorithme

input : la matrice des données **Y**

1. pour tout $i \in 1, \ldots, N$, calculer les coefficients $\hat{\beta}_i$ de la régression parcimonieuse de $\mathbf{y}_i$ sur les autres colonnes de **Y** et former la matrice **B**
2. Construire le graphe G de similarité dont les sommets sont les N points et les poids des arêtes sont donnés par la matrice $\mathbf{W} = |\mathbf{B}| + |\mathbf{B}|^t$
3. Calculer les valeurs propres $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_N$ du Laplacien normalisé de G et

$$\hat{K} = N - \arg \max_{i=1,\ldots,N} (\lambda_i - \lambda_{i+1})$$

4. Appliquer une technique de spectral clustering en utilisant $\hat{K}$ comme nombre estimé des classes
5. ACP pour construire les sous-espaces et débruiter **Y**

output : les sous-espaces et les données débruitées

performances

false discovery : $B_{ij} \neq 0$ et $\mathbf{y}_i$ et $\mathbf{y}_j$ n'appartiennent pas au même sous-espace
true discovery : $B_{ij} \neq 0$ et $\mathbf{y}_i$ et $\mathbf{y}_j$ appartiennent au même sous-espace
2 paramètres importants :

1. proximité entre 2 sous-espaces :

$$aff(S, S') = \sqrt{\text{moyenne des carrés des corrélations canoniques}}$$

2. densité d'un sous-espace :

$$\rho_k = \frac{N_k}{d_k}$$

la régression LASSO

Least Absolute Shrinkage and Selection Operator

$$\min_{\beta \in \mathbb{R}^N} \frac{1}{2} \|\mathbf{y}_i - \mathbf{Y}\beta\|_2^2 + \lambda \|\beta\|_1$$

problème : choix de $\lambda \Rightarrow$ peu de "false discoveries" et beaucoup de "true discoveries"
λ dépend de la dimension d du sous-espace
λ de l'ordre de $1/\sqrt{d}$

régression LASSO en deux étapes

1. résoudre $\beta^* = \arg\min \|\beta\|_1$ vérifiant $\|\mathbf{y}_i - \mathbf{Y}\beta\|_2^2 \leqslant \tau$ et $\beta_i = 0$
2. prendre $\lambda = f(\|\beta^*\|_1)$
3. résoudre $\hat{\beta} = \arg\min \frac{1}{2}\|\mathbf{y}_i - \mathbf{Y}\beta\|_2 + \lambda\|\beta\|_1 0$

Résultats théoriques

2 conditions :

$$\max_{l:l\neq k} aff(S_k, S_l) \leqslant \frac{\kappa_0}{\log N} - \sigma\sqrt{\frac{d_k}{2p\log N}} \text{ et } \rho_k \geqslant \rho*$$

Th1 : Si S_k vérifie les 2 conditions et σ suffisamment petit, alors la k-ème colonne de B ne contient pas de false discoveries avec une forte probabilité pour le choix $\tau = 2\sigma$ et $f(t) \geqslant .707\sigma t^{-1}$

Th2 : sous les mêmes conditions que le théorème 1 et ajoutant la condition$f(t) \leqslant \alpha_0 t^{-1}$, avec une forte probabilité, le nombre de true discoveries dépasse

$$c_0 \frac{d(i)}{\log(\rho(i))}$$