Projection, Discrimination et Classification

Massih-Réza Amini

Techniques d'Analyse de Données et Théorie de l'Information
Master M2 IAD – Parcours Recherche
amini@poleia.lip6.fr

http://www-connex.lip6.fr/~amini

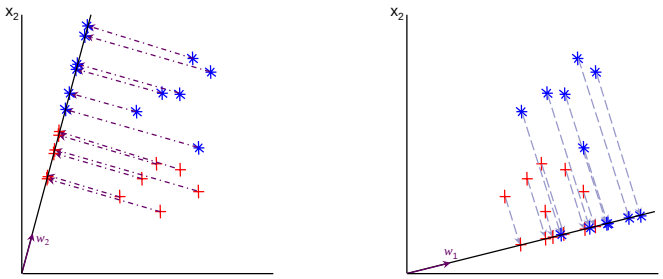
Plan

- Analyse Factorielle Discriminante de Fisher (AFD)
- Régression linéaire au sens des moindre carrées
- Analyse Discriminante Linéaire (ADL)
- Lien entre AFD, ADL et RMC
  - Synthèse sur AFD, ADL, RMC et RL
  - Avantages et Inconvénients de ADL
- Régression Logistique
- Exemple d'application: la tâche de Recherche d'Information
  - Recherche avec un codage binaire,
  - Vers une meilleure représentation,
  - L'information de classe avec l'AF améliore encore la représentation

Analyse discriminante de Fisher

- Pb de dimensionnalité : beaucoup de méthodes / techniques (en classif.) sont inappropriées ou inapplicables en grande dimension

⇒ On projette sur un sous-espace
- Perte possible d'informations
- Trouver la projection optimale à ce sens

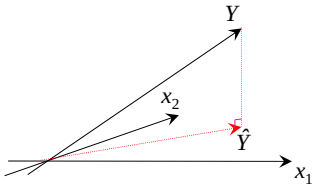


Analyse Discriminante de Fisher (2)

- Critère d'optimisation : $J(w) = \frac{w^t.S_B.w}{w^t.S_W.w}$
- Solution vérifie : $S_B.w = \lambda.S_W.w$
- Solution $w \propto S_W^{-1}.(m_1 - m_2)$

Interprétation géométrique

- La solution de la régression $\hat{B}$ vérifie $X^t\left(Y - X\hat{B}\right) = X^t\left(Y - \hat{Y}\right) = 0$



- La réponse du modèle, $\hat{Y}$ est la projection orthogonale de Y sur l'espace des données.

$$\hat{Y} = X\hat{B} = X\left(X^t X\right)^{-1} X^t Y$$

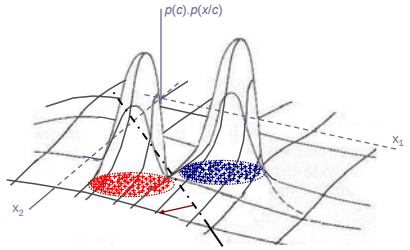
Matrice de projection, vérifie la propriété d'idempotence

Analyse Discriminante Linéaire

- But : Discrimination
- ADL dans le cas de populations normales avec une matrice de covariance commune pour les différentes groupes

$$\log \frac{p(c=k \mid x)}{p(c=l \mid x)} = x^t.\Sigma^{-1}(\mu_k - \mu_l) - \frac{1}{2}(\mu_k + \mu_l)^t \Sigma^{-1}(\mu_k - \mu_l) + \log \frac{\pi_k}{\pi_l}$$

- Règle de décision dans le cas bi-classe $f(x) = x^t.\beta + \beta_0 \begin{cases} > 0, & \text{si } x \in C_1; \\ < 0, & \text{si } x \in C_2. \end{cases}$



Estimation des paramètres du modèle linéaire

- On suppose que chaque exemple x_i appartient à une et une seule classe.

$$\forall x_i \in C_k, \text{ on lui associe un vecteur indicateur de classe } t_i = (t_{1i}, ..., t_{ci})$$

$$x_i \in C_k \Leftrightarrow t_{ki} = 1 \text{ et } \forall h \neq k, t_{hi} = 0$$

- Critère d'optimisation : logarithme de la vraisemblance complète des données (ou la vraisemblance classifiante)

$$V_c = \sum_{i=1}^{n}\sum_{k=1}^{c} t_{ki}.\log\left[p(x_i, c=k)\right] = \sum_{i=1}^{n}\sum_{k=1}^{c} t_{ki}.\log\left[p(c=k).p(x_i \mid c=k)\right]$$

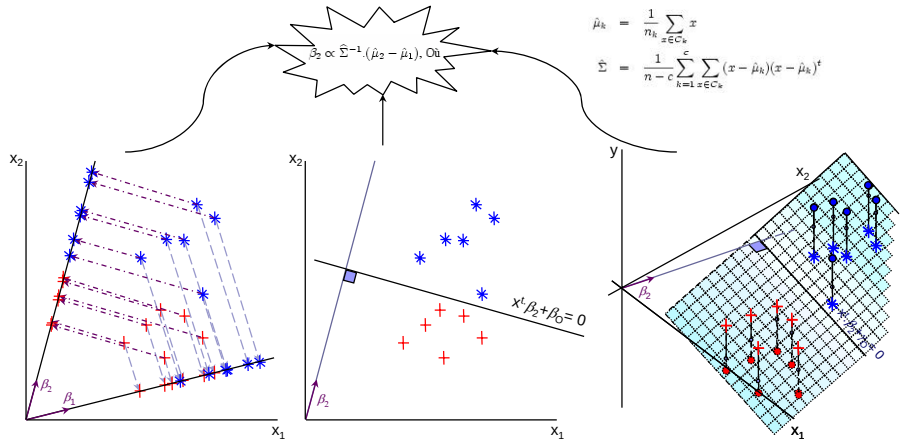
- Dans le cas normal les paramètres du modèle qui maximise ce critère sont

$$p(c=k) = \frac{n_k}{n}, \text{ où } n_k \text{ est le cardinal de la classe } C_k$$

$$\hat{\mu}_k = \frac{1}{n_k}\sum_{x \in C_k} x$$

$$\hat{\Sigma} = \frac{1}{n-c}\sum_{k=1}^{c}\sum_{x \in C_k}(x - \hat{\mu}_k)(x - \hat{\mu}_k)^t$$

Lien entre AF, ADL et RMC



Analyse Factorielle Discriminante — ***Projection***

Analyse Linéaire Discriminante — ***Discrimination***

Régression Linéaire au Moindre Carrées — ***Classification***

Lien entre AF, ADL et RMC

- AFD ∝ RMC
  - Si n_1 et n_2 sont les cardinaux des 2 classes et que les désirées sont codées par $(-1)^k n / n_k$ pour, $k \in \{1,2\}$
  - Les paramètres β qui minimise EMC satisfont l'équation

$$\left[S_W + \frac{n_1 n_2}{n} S_B\right].\beta = n(m_2 - m_1) \Rightarrow \beta \propto S_W^{-1}(m_2 - m_1)$$

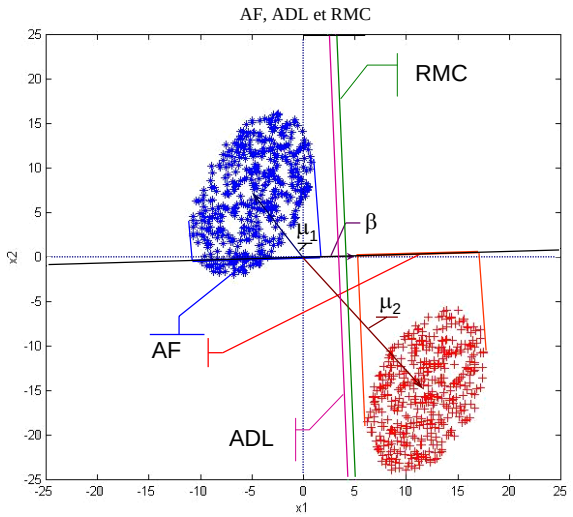
- ADL ∝ RMC
  - On remplace le problème d'inégalités dans le cas ADL

$$f(x) = x^t.\beta + \beta_0 \begin{cases} > 0, & \text{si } x \in C_1; \\ < 0, & \text{si } x \in C_2. \end{cases}$$

  - Par un problème d'égalités dans le cas RMC

$$f(x) = x^t.\beta + \gamma_0 \begin{cases} = 1, & \text{si } x \in C_1; \\ = -1, & \text{si } x \in C_2. \end{cases}$$

AF, LDA et LSR sur un exemple jouet



$$\pi_1 = \pi_2 = \frac{1}{2}$$

$$\mu_1 = \begin{bmatrix} -5 \\ 7 \end{bmatrix}, \quad \mu_2 = \begin{bmatrix} 12 \\ -15 \end{bmatrix}$$

$$\Sigma_1 = \Sigma_2 = \Sigma = \begin{bmatrix} 0.062 & -0.1376 \\ -0.1376 & 1.6169 \end{bmatrix}$$

$$\beta \propto \Sigma^{-1}.(\mu_2 - \mu_1) = \begin{bmatrix} 300.81 \\ 11.99 \end{bmatrix}$$

ADL

$$x^t.\Sigma^{-1}.(\mu_2 - \mu_1) - \frac{1}{2}(\mu_1 + \mu_2)^t.\Sigma^{-1}.(\mu_2 - \mu_1) = 0$$
$$300.81 \times x_1 + 11.99 \times x_2 - 1004.9 = 0$$

RMC

$$x^t.\beta + \beta_0 = 0$$
$$0.248 \times x_1 + 0.0099 \times x_2 - 1.0053 = 0$$

Régression Logistique

- On modélise le logarithme du rapport des densités conditionnelles de classes par des fonctions linéaires. Pour le cas bi-classes :

$$\log \frac{p(x \mid c=1)}{p(x \mid c=2)} = x^t.\beta + \beta_0$$

- Propriétés
  - Les probabilités a posteriori ont une forme simple (logistique)

$$p(c=1 \mid x) = \frac{e^{(x^t.\beta+\beta_0)}}{1+e^{(x^t.\beta+\beta_0)}} = \frac{1}{1+e^{-(x^t.\beta+\beta_0)}}$$
$$p(c=2 \mid x) = \frac{1}{1+e^{(x^t.\beta+\beta_0)}}$$

  - Une grande variété de familles de distributions vérifient l'hypothèse de départ.
- Critère d'optimisation : vraisemblance classifiante

$$V_c = \sum_{i=1}^{n}\sum_{k=1}^{c} t_{ki}.\log[p(x_i, c=k)] = \sum_{i=1}^{n}\sum_{k=1}^{c} t_{ki}.\log[p(x_i).p(c=k \mid x_i)]$$

Régression Logistique (2)

- Critère d'optimisation équivalente (cas bi-classe)

$$L'_c(B) = \sum_{i=1}^{n}[y_i.\log p(c=1 \mid x_i) + (1-y_i).\log(1-p(c=1 \mid x_i))], \text{ où } y_i = t_{1i}$$

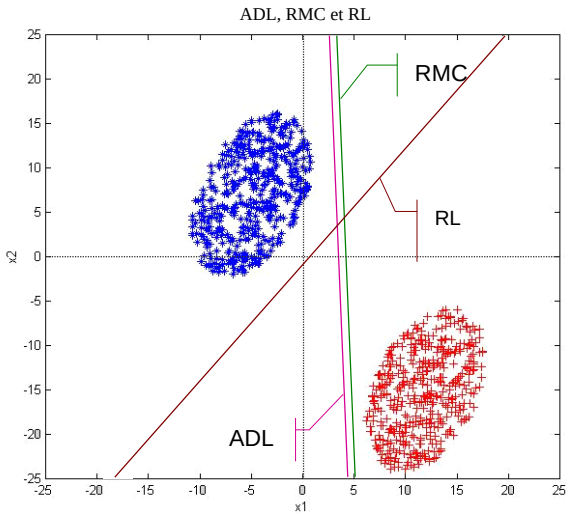
- Les dérivées partielles d'ordre 1 et 2 de L'_c en fonction de B sont :

En posant p le n-vecteur des probabilités a posteriori de classes estimées, et W la matrice diagonale $n \times n$ ayant $p(c=1 \mid x_i).(1-p(c=1 \mid x_i))$ comme le $i^{ième}$ élèment diagonal

$$\frac{\partial L'_c(B)}{\partial B} = \sum_{i=1}^{n} x_i.(y_i - p(c=1 \mid x_i)) = X^t.(Y-p)$$
$$\frac{\partial^2 L'_c(B)}{\partial B.\partial B^t} = -\sum_{i=1}^{n} x_i.x_i^t.p(c=1 \mid x_i).(1-p(c=1 \mid x_i)) = -X^t.W.X$$

- Solution par approximation numérique et pas de solution exacte

ADL, RMC et RL sur l'exemple *jouet*



ADL

$$x^t.\Sigma^{-1}.(\mu_2-\mu_1)-\frac{1}{2}(\mu_1+\mu_2)^t.\Sigma^{-1}.(\mu_2-\mu_1) = 0$$
$$300.81\times x_1+11.99\times x_2-1004.9 = 0$$

RMC

$$x^t.\beta+\beta_0 = 0$$
$$0.248\times x_1+0.0099\times x_2-1.0053 = 0$$

RL

Initialiser les paramètres $B^{(0)}=(B_j^{(0)})_{j=0,\dots,d}$ du modèle logistique $C_{B^{(p)}}$ sur $[0,1]$
Pour $p\geq 0$, *itérer jusqu'à la convergence de* L'_c

1. Calculer la sortie du modèle $C_{B^{(p)}}$, estimant les probabilités à posteriori de classes des exmeples :

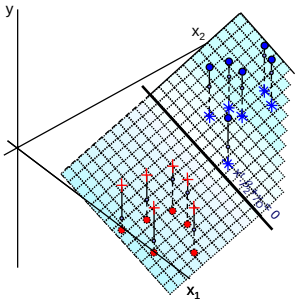
$$\forall x,\ p(c=1\mid x) = \frac{1}{1+exp(\beta_0^{(p)}+\beta^{t(p)}.x)}$$
$$p(c=2\mid x) = 1-p(c=1\mid x)$$

2. Estimer les nouveaux paramtres $B^{(p+1)}$ maximisant $L'_c(B^{(p)})$

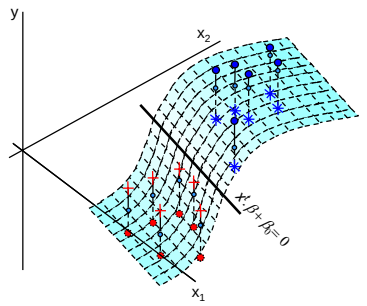
$$B^{(p+1)} \leftarrow B^{(p)}-\left(\frac{\partial^2 L'_c(B^{(p)})}{\partial B\partial B^t}\right)^{-1}.\frac{\partial L'_c(B^{(p)})}{\partial B}$$

$$x^t.\beta+\beta_0 = 0$$
$$1.2098\times x_1-0.9138\times x_2-0.7688 = 0$$

RMC et RL : Deux régresseurs linéaires différentes



Régression Linéaire au Moindre Carrées



Régression Logistique

RL ou ADL ?

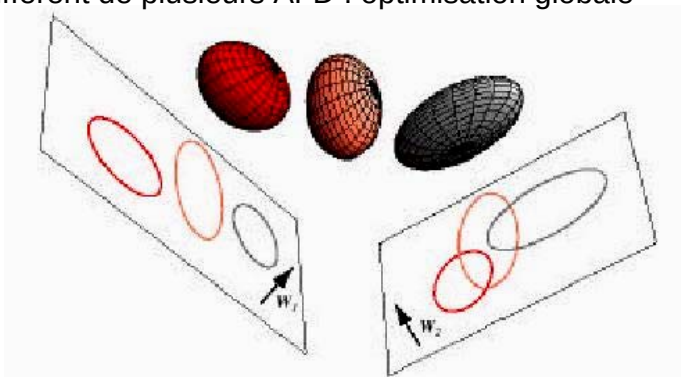
- Les densités gaussiennes avec une même matrice de covariance, vérifient l'hypothèse des modèles logistiques.

$$\log\frac{p(x\mid c=1)}{p(x\mid c=2)} = x^t.\Sigma^{-1}.(\mu_1-\mu_2)-\frac{1}{2}(\mu_1+\mu_2)^t.\Sigma^{-1}.(\mu_1-\mu_2)$$

- Les probabilités a posteriori de classes dans les deux cas (RL et ADL) ont la même forme logistique, mais les coefficients pour ces deux modèles sont estimés de façon différents.
  - La régression logistique ne fait pas d'hypothèse sur la densité marginale des points X et trouve les coefficients de $p(C\ /\ X)$ en maximisant la forme équivalent de la vraisemblance classifiante, L'_c.
  - Avec l'ADL, les paramètres du modèle sont estimées en maximisant la forme complète de la vraisemblance classifiante. Ici, la densité marginale $p(X)$ joue un rôle dans l'estimation de ces paramètres.

Analyse Discriminante Multiple

- *d* à *c*-1 dimension avec *c* le nombre de classes
- Généralisation de Fisher
- *w* devient une matrice W [*d* x (*c*-1)]
- Différent de plusieurs AFD : optimisation globale



Analyse Discriminante Multiple (2)

- *p* à *c*-1 dimension avec *c* le nombre de classes
- Généralisation de Fisher
- *w* devient une matrice *W* [*p* x (*c*-1)]
- Pour une population centrée

$$m = \frac{1}{n}\sum_x x = 0$$

$$m_k = \frac{1}{n_k}\sum_{x \in C_k} x$$

- La matrice de dispersion inter-classe $S_B = \frac{1}{n}\sum_{k=1}^{c} n_k.m_k.m_k^t$
  variance des moyennes
- La matrice de dispersion intra-classe $S_W = \frac{1}{n}\sum_{k=1}^{c}\sum_{x \in C_k}(x - m_k).(x - m_k)^t$
  moyenne des variances

Analyse Discriminante Multiple (3)

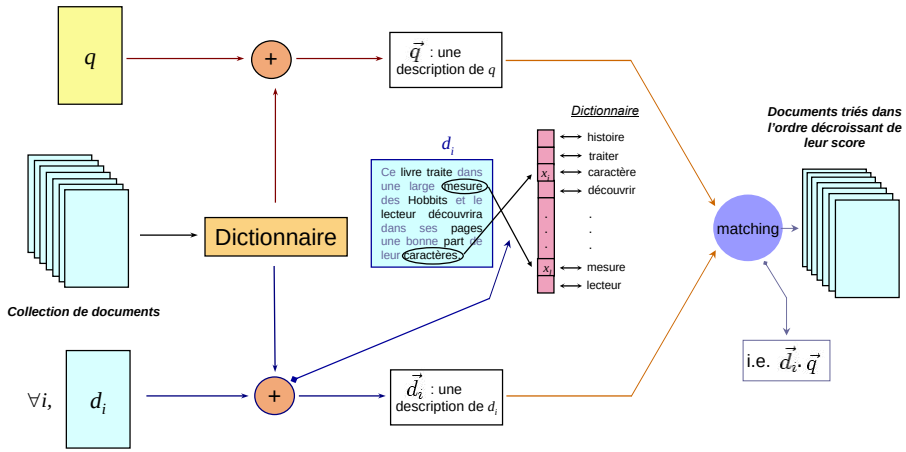
- X, la matrice $n \times p$ des données,
- Y, la matrice $n \times c$ des classes,
- $V = \frac{1}{n}.X^t.X$, la variance totale
- $D = \frac{1}{n}.Y^t.Y = diag(\frac{n_1}{n}, ..., \frac{n_c}{n})$
- $\Sigma = \frac{1}{n}.X^t.Y = Tr\left[Vect(\frac{1}{n}.\sum_{x \in C_k} x)_{k \in \{1,...,c\}}\right]$
- $M = D^{-1}.\Sigma = Tr\left[Vect(\frac{1}{n_k}.\sum_{x \in C_k} x)_{k \in \{1,...,c\}}\right] = Tr\left[Vect(m_k)_{k \in \{1,...,c\}}\right]$

Posons, $P_Y = Y.(Y^t.Y)^{-1}.Y^t$ le projecteur sur l'espace des colonnes de Y

- $S_B = \frac{1}{n}.(P_Y.X)^t(P_Y.X) = \Sigma^t.D^{-1}.\Sigma = M^t.D.M$
- $S_W = \frac{1}{n}\left[((I_n - P_Y).X)^t((I_n - P_Y).X)\right] = V - S_B$

- Critère : Chercher W qui rend stationnaire $W^t.S_B.W$ *i.e.* $\lim_{n\to\infty}(W^t.S_B.W)^n$ est finie, sous la contrainte $W^t.S_W.W = I_{c-1}$

Exemple d'application: Recherche de l'Information



Exemple Jouet

- 18 documents codés sur un dictionnaire de 32 termes. Les documents sont classés suivant trois catégories : ***Business Recession*** (classe *A*), ***Tectonics*** (classe B) et ***Mental disorder*** (classe *C*).

Classe A	Recession	A1	business, depression, tax, unemployment
	Stagflation	A2	business, economy, market
	Business Indicator	A3	business, market, production
	Business Cycle	A4	depression, liquidation, prosperity, recovery
	Unemployment	A5	compensation, unemployment, welfare
	Inflation	A6	business, market, price
	Welfare	A7	benefit, compensation, unemployment
Classe B	Continental Drift	B1	basin, fault, submergence
	Plate Tectonics	B2	depression, fault
	Great Basin	B3	basin, drainage, valley
	Valley	B4	depression, drainage, erosion
	Lake Formation	B5	basin, drainage, volcano
Classe C	Psychosis	C1	counseling, illness, mental
	Alcoholism	C2	counseling, depression, emotion
	Mental Health	C3	depression, rehabilitation, treatment
	Depression Disorder	C4	disturbance, drug, illness
	Manic Depression	C5	distractibility, illness, talkativeness
	Mental Treatment	C6	counseling, drug, psychotherapy

Tâche de recherche

- But : Rechercher les documents les plus pertinents par rapport à une requête *q* et les triés dans l'ordre décroissant de leur pertinence.
- La requête *q* = {business, depression}

Recession	A1	business, depression, tax, unemployment
Stagflation	A2	business, economy, market
Business Indicator	A3	business, market, production
Business Cycle	A4	depression, liquidation, prosperity, recovery
Unemployment	A5	compensation, unemployment, welfare
Inflation	A6	business, market, price
Welfare	A7	benefit, compensation, unemployment
Continental Drift	B1	basin, fault, submergence
Plate Tectonics	B2	depression, fault
Great Basin	B3	basin, drainage, valley
Valley	B4	depression, drainage, erosion
Lake Formation	B5	basin, drainage, volcano
Psychosis	C1	counseling, illness, mental
Alcoholism	C2	counseling, depression, emotion
Mental Health	C3	depression, rehabilitation, treatment
Depression Disorder	C4	disturbance, drug, illness
Manic Depression	C5	distractibility, illness, talkativeness
Mental Treatment	C6	counseling, drug, psychotherapy

Recherche d'Information avec un codage binaire

	1	3	3	9	10	3	10	10	1	10	3	10	10	3	3	10	10	10	
	Documents																		
	A1	A2	A3	A4		A6			B2		B4			C2	C3				
basin	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0
benefit	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
requête *q* → business	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
compensation	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
counseling	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	
requête *q* → depression	1	0	0	1	0	0	0	0	1	0	1	0	0	1	1	0	0	0	
distractibility	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0
disturbance	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0		
drainage	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	
drug	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	
economy	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
emotion	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	
erosion	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	
fault	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	
illness	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	
liquidation	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
market	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
mental	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	
price	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	
production	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
prosperity	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
psychotherapy	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	
recovery	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
rehabilitation	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
submergence	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	
talkativeness	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	
tax	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
treatment	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	
unemployment	1	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	
valley	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	
volcano	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	
welfare	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Codage insuffisant

- Codage binaire est insuffisant pour capturer une information de contenue

Classe A	Recession	A1	business, depression, tax, unemployment
	Stagflation	A2	business, economy, market
	Business Indicator	A3	business, market, production
	Business Cycle	A4	depression, liquidation, prosperity, recovery
	Unemployment	A5	compensation, unemployment, welfare
	Inflation	A6	business, market, price
	Welfare	A7	benefit, compensation, unemployment
Classe B	Continental Drift	B1	basin, fault, submergence
	Plate Tectonics	B2	depression, fault
	Great Basin	B3	basin, drainage, valley
	Valley	B4	depression, drainage, erosion
	Lake Formation	B5	basin, drainage, volcano
Classe C	Psychosis	C1	counseling, illness, mental
	Alcoholism	C2	counseling, depression, emotion
	Mental Health	C3	depression, rehabilitation, treatment
	Depression Disorder	C4	disturbance, drug, illness
	Manic Depression	C5	distractibility, illness, talkativeness
	Mental Treatment	C6	counseling, drug, psychotherapy

Polysémie

Synonymie

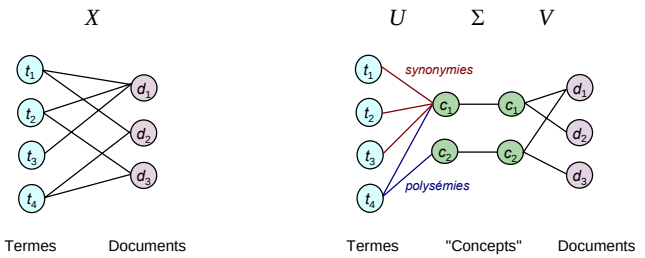
Codage Insuffisant (2)

- But : Rechercher les documents les plus pertinents par rapport à une requête *q* et les triés dans l'ordre décroissant de leur pertinence.
- La requête *q* = {business, depression}

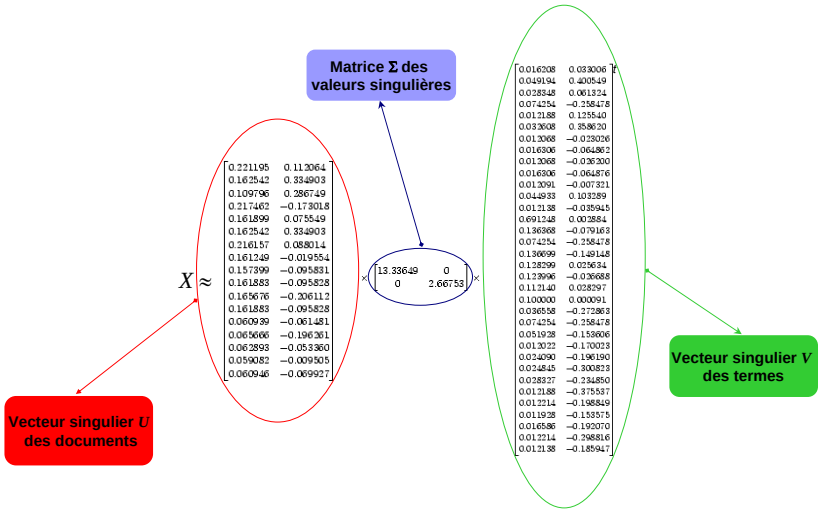
Recession	A1	business, depression, tax, unemployment
Stagflation	A2	business, economy, market
Business Indicator	A3	business, market, production
Business Cycle	A4	depression, liquidation, prosperity, recovery
Unemployment	A5	compensation, unemployment, welfare
Inflation	A6	business, market, price
Welfare	A7	benefit, compensation, unemployment
Continental Drift	B1	basin, fault, submergence
Plate Tectonics	B2	depression, fault
Great Basin	B3	basin, drainage, valley
Valley	B4	depression, drainage, erosion
Lake Formation	B5	basin, drainage, volcano
Psychosis	C1	counseling, illness, mental
Alcoholism	C2	counseling, depression, emotion
Mental Health	C3	depression, rehabilitation, treatment
Depression Disorder	C4	disturbance, drug, illness
Manic Depression	C5	distractibility, illness, talkativeness
Mental Treatment	C6	counseling, drug, psychotherapy

Représentation plus riche: Latent Semantic Indexing

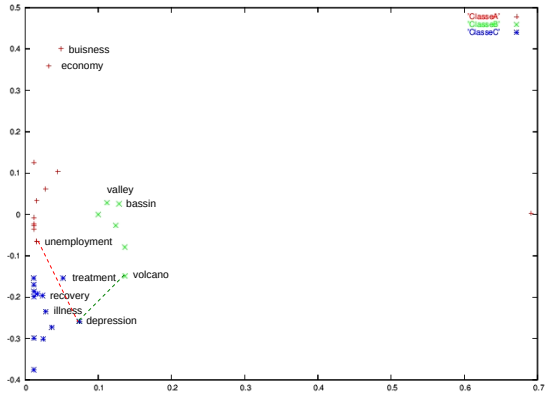
- Le LSI est la décomposition en valeurs singulières d'une matrice terme-document $X = U.\Sigma.V^t$



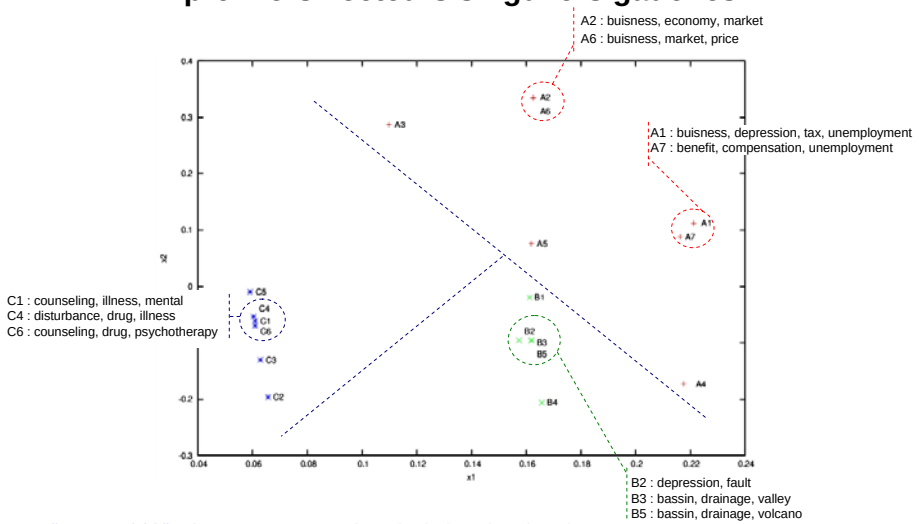
Décomposition de la matrice des données en valeurs et vecteurs singuliers



Représentation des termes sur la base des 2 premiers vecteurs singuliers droites



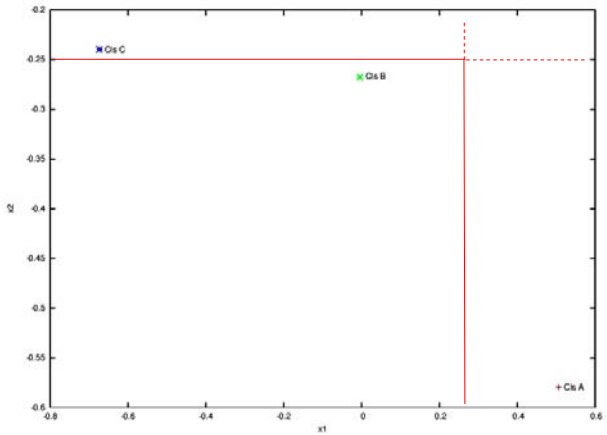
Représentation des documents sur la base des 2 premiers vecteurs singuliers gauches



Liste des documents retournés

Pertinence par rapport à q			
RI classique		LSI	
Docs.	Score	Docs.	Score
A1	1.0	A1	1.0
B2	1.0	A7	.99
A2	.57	A5	.99
A3	.57	C3	.77
A6	.57	B2	.76
B2	.57	A4	.75
B4	.57	C2	.74
C2	.57	B4	.71
C3	.57	A2	.65
A5	0	A3	.65
A7	0	A6	.52
B1	0	C6	.52
B3	0	C1	.50
B5	0	B1	.43
C1	0	B3	.40
C4	0	B5	.40
C5	0	C4	.34
C6	0	C5	.31

Représentation des documents sur la base des 2 premiers vecteurs propres de Fisher



$$\sum_{i=1}^{18} x_i = 0$$

$$m_k = \frac{1}{n_k} \sum_{x \in C_k} x$$

$$S_B = \frac{1}{18} \sum_{k=1}^{3} n_k . m_k . m_k^t$$

$$S_W = \frac{1}{18} \sum_{k=1}^{3} \sum_{x \in C_k} (x - m_k)(x - m_k)^t$$

S_w est non-singulière, on cherche les (c-1=2) valeurs et vecteurs propres de

$$(S_W + \epsilon . I)^{-1} . S_B \text{ avec } \epsilon = 10.e^{-4}$$

$\forall x$, le projetté de x sur les vecteurs propres V_i : $x^t . V_i$