Analyse implicative et modèle de Rasch

Yvonnick Noël

Université Rennes 2, LP3C

MODEVAIIA, 21 juin 2017

Sommaire

Motivation

- Nous nous intéressons à la mesure de **force associative** entre deux variables binaires.
- Les indices d'associations connaissent un vif regain d'intérêt avec le développement du Machine Learning : devant le Big Data, on cherche des modes d'**extraction automatique** de structures associatives.
- Il existe un grand nombre d'indices associatifs (Huynh, 2007, en recense 36 principaux), la plupart basés sur la notion de **co-fréquence**.
- Cela peut être la couche de base d'**analyses plus complexes** : en clusters, en dimensions ou en graphe (lien au Rasch à expliciter).

Notations

A/B	b	$\bar{b}$	Somme
a	n_{ab}	$n_{a\bar{b}}$	n_a
$\bar{a}$	$n_{\bar{a}b}$	$n_{\bar{a}\bar{b}}$	$n_{\bar{a}}$
Somme	n_b	$n_{\bar{b}}$	N

A/B	b	$\bar{b}$	Somme
a	π_{ab}	$\pi_{a\bar{b}}$	π_a
$\bar{a}$	$\pi_{\bar{a}b}$	$\pi_{\bar{a}\bar{b}}$	$\pi_{\bar{a}}$
Somme	π_b	$\pi_{\bar{b}}$	1

- La **co-fréquence** (ou fréquence conjointe) pure par exemple est calculable comme :

$$f_{ab} = \frac{n_{ab}}{N}$$

et associée dans le raisonnement statistique à une **probabilité conjointe** inconnue π_{ab}.

La corrélation

- Le coefficient ϕ :

$$\phi = \frac{n_{ab}n_{\bar{a}\bar{b}} - n_{a\bar{b}}n_{\bar{a}b}}{\sqrt{n_a n_{\bar{a}} n_b n_{\bar{b}}}}$$

n'est autre que la **corrélation de Pearson** sur variables binaires.

- Il est lié au χ^2 de contingence par la relation :

$$\phi^2 = \frac{\chi^2}{N}$$

Le rapport de cote

- Le rapport de cote :

$$o_{AB} = \frac{f_{b|a}/f_{\bar{b}|a}}{f_{b|\bar{a}}/f_{\bar{b}|\bar{a}}} = \frac{n_{ab}/n_{a\bar{b}}}{n_{\bar{a}b}/n_{\bar{a}\bar{b}}} = \frac{n_{ab}n_{\bar{a}\bar{b}}}{n_{\bar{a}b}n_{a\bar{b}}}$$

mesure comment la disproportion b et $\bar{b}$ est différente selon qu'on dans le cas a ou $\bar{a}$.

A/B	b	$\bar{b}$	Somme
a	n_{ab}	$n_{a\bar{b}}$	n_a
$\bar{a}$	$n_{\bar{a}b}$	$n_{\bar{a}\bar{b}}$	$n_{\bar{a}}$
Somme	n_b	$n_{\bar{b}}$	N

L'indice de Jaccard

- L'indice de Jaccard :

$$j_{ab} = \frac{n_{ab}}{n_{ab} + n_{\bar{a}b} + n_{a\bar{b}}}$$

mesure la co-fréquence sur les seuls cas où **au moins A ou B** s'est produit.

A/B	b	$\bar{b}$	Somme
a	n_{ab}	$n_{a\bar{b}}$	n_a
$\bar{a}$	$n_{\bar{a}b}$	$n_{\bar{a}\bar{b}}$	$n_{\bar{a}}$
Somme	n_b	$n_{\bar{b}}$	N

Indices dissymétriques

- Le coefficient ϕ, le rapport de cote ou l'indice de Jaccard sont **symétriques** : leur valeur reste inchangée si on permute les symboles a et b.
- Nous sommes intéressés à trouver des indices de dépendance dissymétriques, pour rendre compte du fait que dans des échelles psychologiques, la réussite sur un item peut impliquer la réussite sur un autre, mais **pas nécessairement l'inverse**.
- A la suite de Gras (2008), nous allons appeler **indices d'implication** de tels indices.

L'indice d'homogénéité de Loevinger

- L'indice de Loevinger (1947) se calcule sur une paire ordonnée d'items (A est le plus difficile) :

$$H = \frac{f_{ab} - f_a f_b}{f_a - f_a f_b} = \frac{f_a - f_{a\bar{b}} - f_a f_b}{f_a - f_a f_b} = 1 - \frac{f_{a\bar{b}}}{f_a f_{\bar{b}}}$$

  mesure l'écart à l'indépendance uniquement dans la case des co-fréquences, conformes à l'implication $A \to B$, rapporté à sa valeur maximale (quand la fréquence $f_{a\bar{b}}$ des contre-exemples est nulle).
- Propriétés : il est dissymétrique, égal à 0 dans le cas de l'indépendance, et égal à 1 quand il n'y a aucun contre-exemple à la règle (il peut aussi être négatif).
- Inconvénient connu : en cas de difficulté égale des items ($f_a = f_b$) et $n_{\bar{a}b} = 0$, il peut aussi valoir 1.

Principe

- Dans le tableau de distribution conjointe :

A/B	b	$\bar{b}$	Somme
a	π_{ab}	$\pi_{a\bar{b}}$	π_a
$\bar{a}$	$\pi_{\bar{a}b}$	$\pi_{\bar{a}\bar{b}}$	$\pi_{\bar{a}}$
Somme	π_b	$\pi_{\bar{b}}$	1

  nous attendons $\pi_{a\bar{b}} = 0$ en cas d'**implication parfaite** $A \to B$.
- Cette exigence ne prend pas en compte la multiplicité des causes possibles et nous définissons l'implication statistique comme une relation dissymétrique probabiliste où les **contre-exemples sont rares** : $\pi_{a\bar{b}} \ll 1$.

Modèle

- On se donne la statistique $\nu_{a\bar{b}}$, **comptage des contre-exemples** à l'implication $A \to B$.
- Si les événements sont comptés à partir d'un flux à débit temporel constant, pendant un intervalle fixe d'observation, on peut utiliser un **modèle Poissonien** de comptage espéré μ :

$$P(\nu_{a\bar{b}} = k) = \frac{\mu^k}{k!} e^{-\mu}$$

- Dans la cas de l'indépendance, sur un total d'observations fixé N, le **comptage espéré** est $\mu = N\pi_a\pi_{\bar{b}}$.
- Par les propriétés de la Poisson, la **variance** de ce comptage est égal à la même valeur.

Statistique

- Gras (1979) a proposé la **statistique centrée-réduite** suivante :

$$Q_{a\bar{b}} = \frac{\nu_{a\bar{b}} - N\pi_a\pi_{\bar{b}}}{\sqrt{N\pi_a\pi_{\bar{b}}}}$$

- Sur **données observées**, en estimant π_a et $\pi_{\bar{b}}$ par $\frac{n_a}{N}$ et $\frac{n_{\bar{b}}}{N}$ respectivement, on calcule :

$$q_{a\bar{b}} = \frac{n_{a\bar{b}} - \frac{n_a n_{\bar{b}}}{N}}{\sqrt{\frac{n_a n_{\bar{b}}}{N}}}$$

- Si les effectifs attendus sous le modèle sont au moins de 5, on peut invoquer **l'approximation de la Poisson** par la Normale et considérer cette statistique comme distribuée $N(0, 1)$.

Un exemple

- Bernard & Charron (1996) rapportent une expérience sur la **construction du nombre** chez l'enfant, dans laquelle deux des épreuves consistent à déterminer une quantité par l'intermédiaire d'une fraction qui peut exprimer, soit un rapport **partie-partie** (A), soit un rapport **partie-tout** (B).
- Les réussites et échecs aux deux épreuves sont résumables ainsi :

A/B	b	$\bar{b}$	Somme
a	36	3	39
$\bar{a}$	36	90	126
Somme	72	93	165

Motivation
Indices d'association sur données binaires
L'analyse statistique implicative
Un modèle dimensionnel de l'implication
Conclusions

Description

Observés (n)				Théoriques (n^*)			
A/B	b	$\bar{b}$	Somme	A/B	b	$\bar{b}$	Somme
a	36	3	39	a	17	22	39
$\bar{a}$	36	90	126	$\bar{a}$	55	71	126
Somme	72	93	165	Somme	72	93	165

- On note que 36/39 (92.3%) des élèves qui réussissent l'épreuve Partie-partie réussissent l'épreuve Partie-tout, alors que la réussite globale à cette épreuve n'est pas majoritaire (72/65 soit 43.6%).
- On note que la **réciproque n'est pas vraie** : parmi les 72 qui réussissent l'épreuve Partie-tout, 50% exactement réussissent l'autre.

Motivation
Indices d'association sur données binaires
L'analyse statistique implicative
Un modèle dimensionnel de l'implication
Conclusions

Indice de Loevinger

- L'indice de Loevinger dans le sens $A \to B$ donne :

$$H_{A\to B} = 1 - \frac{n_{a\bar{b}}}{n^*_{a\bar{b}}} = 1 - \frac{3}{22} = 0.863$$

- et dans l'autre sens :

$$H_{B\to A} = 1 - \frac{n_{\bar{a}b}}{n^*_{\bar{a}b}} = 1 - \frac{36}{55} = 0.345$$

- La réussite à l'épreuve A semble davantage permettre de prédire la réussite à l'épreuve B que l'inverse.

Motivation
Indices d'association sur données binaires
L'analyse statistique implicative
Un modèle dimensionnel de l'implication
Conclusions

Statistique Q

- L'indice Q dans le sens $A \to B$ donne :

$$q_{a\bar{b}} = \frac{n_{a\bar{b}} - n^*_{a\bar{b}}}{\sqrt{n^*_{a\bar{b}}}} = \frac{3-22}{\sqrt{22}} = -4.05$$

- et dans l'autre sens :

$$q_{\bar{a}b} = \frac{n_{\bar{a}b} - n^*_{\bar{a}b}}{\sqrt{n^*_{\bar{a}b}}} = \frac{36-55}{\sqrt{55}} = -2.56$$

- A nouveau, la réussite à l'épreuve A semble davantage permettre de prédire la réussite à l'épreuve B que l'inverse, mais la deuxième statistique n'est pas négligeable.

Limites des indices existants

- Ils ne sont pas symétriques dans une relation où l'on attend des relations d'**implication inversée**.
- Ils prennent des valeurs maximum dans des cas **non impliquants** (Loevinger).
- Ils ne permettent pas de penser les relations d'implication dans une **structure dimensionnelle**.
- Ils n'exploitent pas les données disponibles sur la **contraposée** : si $A \to B$ alors $\bar{B} \to \bar{A}$.

Attendu 1

- Dans la distribution conditionnelle de a :

A/B	b	$\bar{b}$	Somme
a	π_{ab}	$\pi_{a\bar{b}}$	π_a
$\bar{a}$	$\pi_{\bar{a}b}$	$\pi_{\bar{a}\bar{b}}$	$\pi_{\bar{a}}$
Somme	π_b	$\pi_{\bar{b}}$	1

nous attendons une **cote favorable** sur b ($o_{b|a} > 1$) si l'implication statistique $A \to B$ existe :

$$o_{b|a} = \frac{\pi_{ab}/\pi_a}{\pi_{a\bar{b}}/\pi_a} = \frac{\pi_{ab}}{\pi_{a\bar{b}}}$$

Attendu 2

- Dans la distribution conditionnelle sur $\bar{b}$:

A/B	b	$\bar{b}$	Somme
a	π_{ab}	$\pi_{a\bar{b}}$	π_a
$\bar{a}$	$\pi_{\bar{a}b}$	$\pi_{\bar{a}\bar{b}}$	$\pi_{\bar{a}}$
Somme	π_b	$\pi_{\bar{b}}$	1

nous attendons une **cote favorable** sur $\bar{a}$ ($o_{\bar{a}|\bar{b}} > 1$) si l'implication statistique $\bar{B} \to \bar{A}$ existe :

$$o_{\bar{a}|\bar{b}} = \frac{\pi_{\bar{a}\bar{b}}/\pi_{\bar{b}}}{\pi_{a\bar{b}}/\pi_{\bar{b}}} = \frac{\pi_{\bar{a}\bar{b}}}{\pi_{a\bar{b}}}$$

Statistique d'implication

- Au final, les deux attendus sont réunis dans la **statistique produit** (indice « iota ») :

$$\iota_{A\to B} = o_{b|a} \times o_{\bar{a}|\bar{b}} = \frac{\pi_{ab}}{\pi_{a\bar{b}}} \times \frac{\pi_{\bar{a}\bar{b}}}{\pi_{a\bar{b}}} = \frac{\pi_{ab}\pi_{\bar{a}\bar{b}}}{\pi^2_{a\bar{b}}}$$

- Propriétés :
  - il varie entre 0 et $+\infty$ (implication parfaite, $\pi_{a\bar{b}} = 0$).
  - Dans ce dernier cas, **l'indice i est infini** (discussion plus loin).
  - Il vaut 1 en situation d'**incertitude totale** ($\pi_{ab} = \pi_{a\bar{b}}$ et $\pi_{\bar{a}\bar{b}} = \pi_{a\bar{b}}$).

Variante

- Pour des raisons qui vont apparaître par la suite, nous considérons la **version log-transformée** :

$$\ln \iota_{A\to B} = \ln\left[\frac{\pi_{ab}\pi_{\bar{a}\bar{b}}}{\pi^2_{a\bar{b}}}\right]$$
$$= \ln \pi_{ab} + \ln \pi_{\bar{a}\bar{b}} - 2\ln \pi_{a\bar{b}}$$

- Sous cette nouvelle forme, l'indice varie de $-\infty$ à $+\infty$, avec la valeur 0 en situation d'incertitude.
- Sur données observées, en estimant les probabilités inconnues par les **fréquences empiriques**, on a :

$$\ln i_{A\to B} = \ln n_{ab} + \ln n_{\bar{a}\bar{b}} - 2\ln n_{a\bar{b}}$$

Exemple

- Sur notre exemple :

Observés (n)			
A/B	b	$\bar{b}$	Somme
a	36	3	39
$\bar{a}$	36	90	126
Somme	72	93	165

on obtient :

$$\ln i_{A\to B} = \ln n_{ab} + \ln n_{\bar{a}\bar{b}} - 2\ln n_{a\bar{b}} = \ln 36 + \ln 90 - 2\ln 3 = 5.886$$
$$\ln i_{B\to A} = \ln n_{ab} + \ln n_{\bar{a}\bar{b}} - 2\ln n_{\bar{a}b} = \ln 36 + \ln 90 - 2\ln 36 = 0.916$$

Notion d'échelle implicative

- Nous nous intéressons dans cette partie au cas où un ensemble d'items peut légitimement être considéré comme une **échelle implicative**.
- Le succès à un item difficile implique le succès à un item plus facile, à condition qu'ils relèvent bien de la **même compétence**.
- Si 3 items de niveaux de difficulté très différents appartiennent à la même échelle implicative, alors le succès sur le plus difficile implique le succès sur les deux autres, mais **plus probablement sur le plus facile**.

Probabilités de succès dans le Rasch

- Pour tout item X_j d'une échelle de Rasch, de difficulté δ_j, les **probabilités** de succès et d'échec sont données par :

$$P(X_j = 1|\theta) = \frac{\exp(\theta - \delta_j)}{1 + \exp(\theta - \delta_j)}$$
$$P(X_j = 0|\theta) = \frac{1}{1 + \exp(\theta - \delta_j)}$$

- Les **probabilités conjointes**, supposant l'indépendance conditionnelle, sont :

X_1/X_2	1	0
1	$\pi_{11} = \frac{\exp(\theta-\delta_1)\exp(\theta-\delta_2)}{[1+\exp(\theta-\delta_1)][1+\exp(\theta-\delta_2)]}$	$\pi_{10} = \frac{\exp(\theta-\delta_1)}{[1+\exp(\theta-\delta_1)][1+\exp(\theta-\delta_2)]}$
0	$\pi_{01} = \frac{\exp(\theta-\delta_2)}{[1+\exp(\theta-\delta_1)][1+\exp(\theta-\delta_2)]}$	$\pi_{00} = \frac{1}{[1+\exp(\theta-\delta_1)][1+\exp(\theta-\delta_2)]}$

Cote conditionnelles et implication

- Selon ce modèle, les cotes conditionnelles pour l'implication et sa contraposée sont données par :

$$\begin{aligned} o_{1.} &= \frac{\pi_{11}}{\pi_{10}} = \exp(\theta - \delta_2) \\ o_{.0} &= \frac{\pi_{00}}{\pi_{10}} = \frac{1}{\exp(\theta - \delta_1)} \end{aligned}$$

  et l'indice ι devient :

$$\iota_{R_1 \to R_2} = o_{1.} \times o_{.0} = \exp(\delta_1 - \delta_2)$$

- Autrement dit :

$$\ln \iota_{R_1 \to R_2} = \delta_1 - \delta_2$$

Implication inverse

- Dans l'autre sens, on a :

$$\begin{aligned} o_{.1} &= \frac{\pi_{11}}{\pi_{01}} = \exp(\theta - \delta_1) \\ o_{0.} &= \frac{\pi_{00}}{\pi_{01}} = \frac{1}{\exp(\theta - \delta_2)} \end{aligned}$$

  et l'indice $\iota_{R_2 \to R_1}$ inverse est :

$$\iota_{R_2 \to R_1} = o_{.1} \times o_{0.} = \exp(\delta_2 - \delta_1) = \frac{1}{\iota}$$

- Autrement dit :

$$\ln \iota_{R_2 \to R_1} = -(\delta_1 - \delta_2) = -\ln \iota_{R_1 \to R_2}$$

Notion de distance implicative

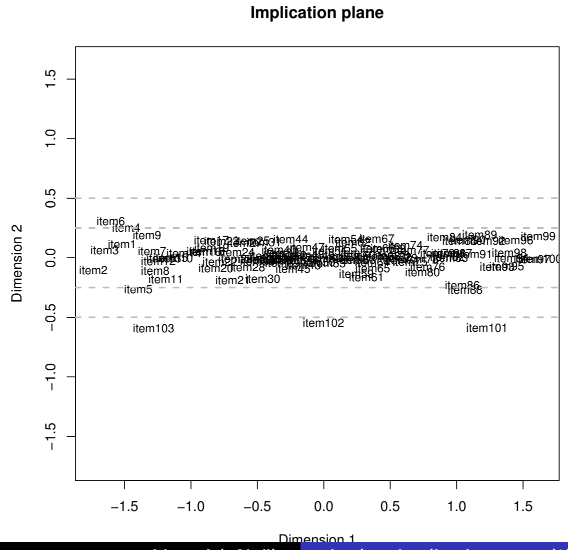
- Dans le cas Rasch, l'implication statistique augmente **avec la distance latente**.
- Sur 3 items ordonnés X_1, X_2 et X_3, dans le cas parfait, on a (**additivité des distances**) : $(\delta_3 - \delta_1) = (\delta_3 - \delta_2) + (\delta_2 - \delta_1)$.
- Sur les indices $|\ln i|$ observés, en valeurs absolues, on devrait pouvoir retrouver une **additivité approximative** sur les distances inter-items d_{ij} qui s'en déduisent :

$$\hat{d}_{13}^{(1)} \approx \hat{d}_{12}^{(1)} + \hat{d}_{23}^{(1)}$$

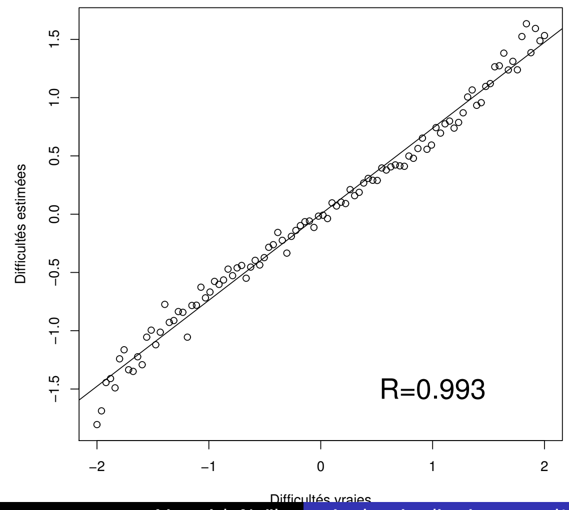
Échelonnement multidimensionnel de Rasch

- Stratégie : on calcule les $|\ln i|$ pour toutes les paires d'items, et on lance un **MDS** sur ces « distances supposées ».
- Si on a une vraie échelle de Rasch, les points items devraient apparaître **approximativement alignés**.
- Les coordonnées en projection sur ces directions d'alignement devraient **estimer les vraies difficultés**.
- Si on a plusieurs échelles de Rasch indépendantes, chaque sous-ensemble devrait avoir son **alignement propre**.

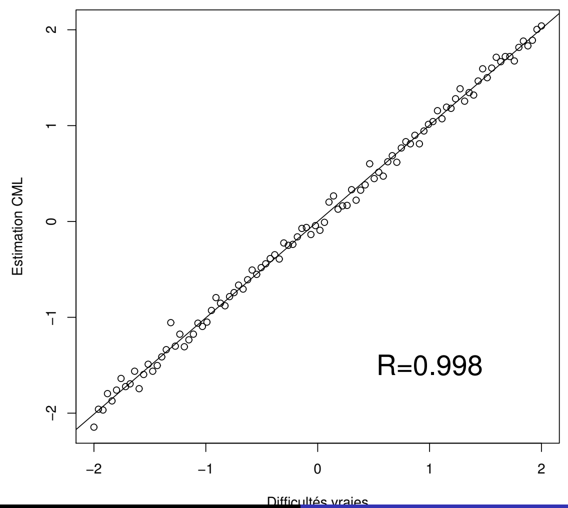
Une échelle de Rasch



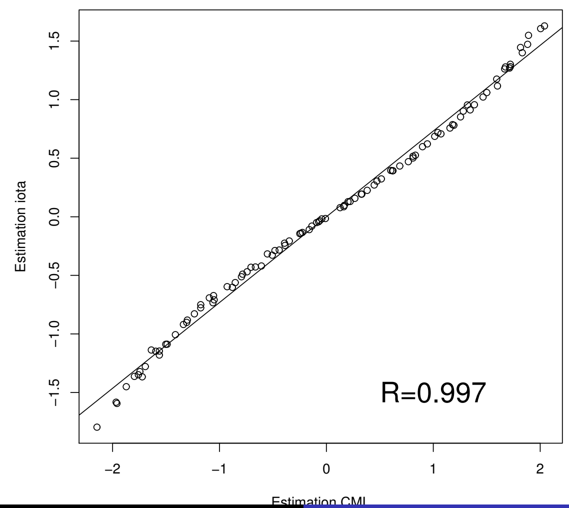
Précision des estimations Iota



Précision des estimations CML (eRm)



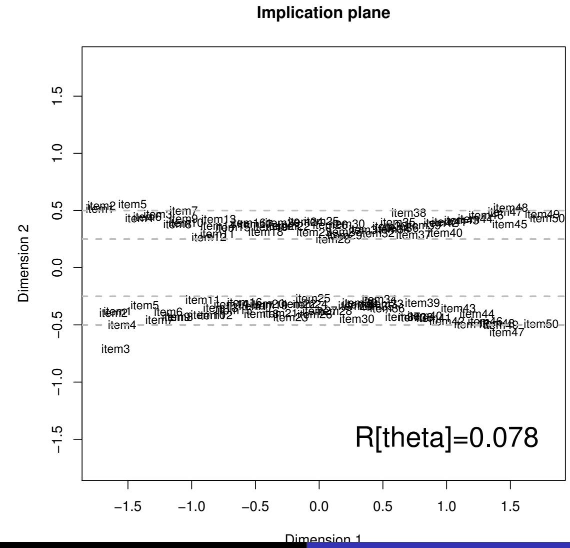
Concordance des estimations CML et Iota



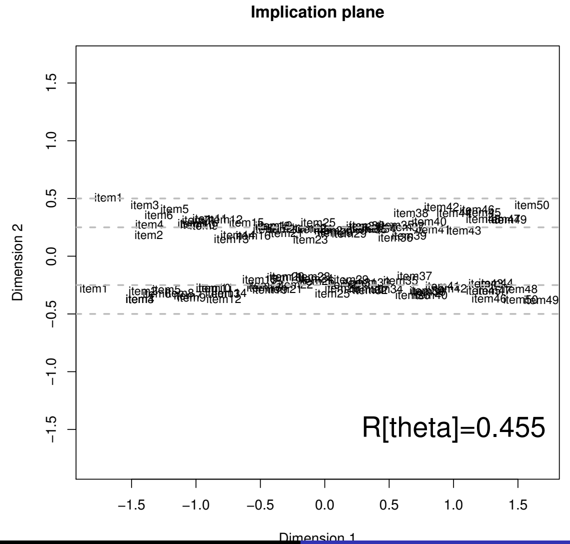
Temps de calcul

- Temps de calcul :
  - Implication Iota : 0.02 seconde.
  - CML : 17 secondes.

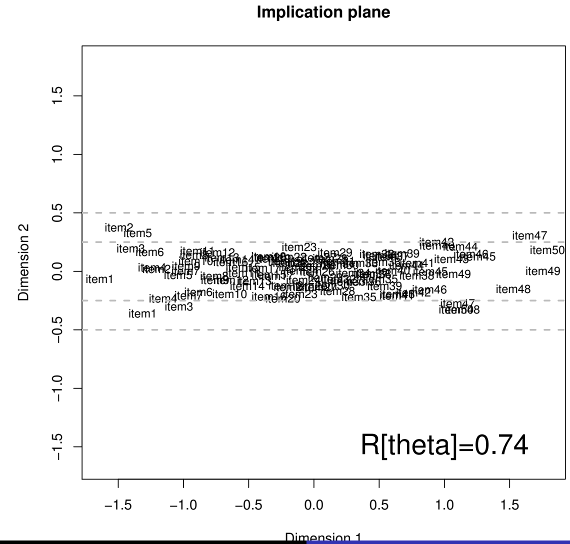
Concordance des estimations CML et Iota



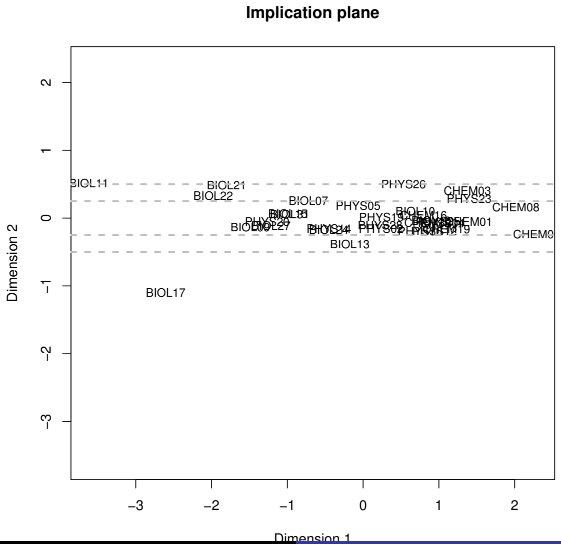
Concordance des estimations CML et Iota



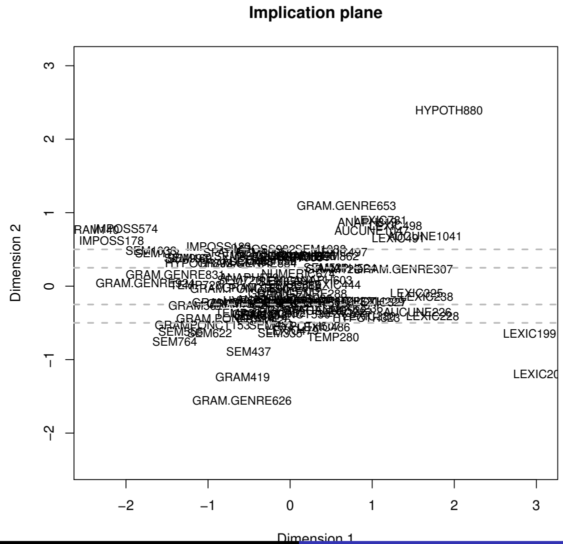
Concordance des estimations CML et Iota



Analyse de tests de sciences



Analyse d'items TACIT



Association, implication, causalité

- On recommande souvent aux étudiants de ne pas confondre **association** (ou corrélation) et **causalité**.
- On a raison... mais on peut aussi étudier statistiquement les relations de **dépendance dissymétriques** (implication).
- L'implication statistique doit cependant être **distinguée de la causalité** simple : l'implication peut être inscrite dans un réseau de dépendances plus complexe (médiations causales).
- L'analyse distancielle sur indices Iota peut être un outil pour explorer graphiquement ces **dépendances en réseau**.

Merci

Merci de votre attention (yvonnick.noel@univ-rennes2.fr)