# Chapitre 2 : Qualité d'un graphe implicatif : variance implicative

Régis Gras*, Jean-Claude Régnier**

*LINA– Ecole Polytechnique de l'Université de Nantes, UMR 6241
La Chantrerie BP 60601 44306 Nantes cedex
regisgra@club-internet.fr
** Université de Lyon - UMR 5191 ICAR
ENS-LSH 15, Parvis René Descartes BP 7000 69342 LYON cedex 07
jean-claude.regnier@univ-lyon2.fr

**Résumé**. Un graphe pondéré, sans cycle, constitue une des représentations d'un ensemble de règles d'association implicative extraites d'un tableau numérique croisant variables et sujets. Le problème de son homogénéité, de sa cohérence et donc de la pertinence des interprétations de l'expert se pose dès lors qu'en Analyse Statistique Implicative (A.S.I.) il est possible de faire varier le seuil de représentation des règles partielles. Nous présentons ici le concept de variance implicative à l'instar du concept classique de variance afin de qualifier l'homogénéité de la représentation. Elle s'appuie sur une métaphore de répulsion vs consistance implicatives mutuelles entre deux variables binaires à partir de leur différence symétrique.

## 1 Introduction

Le texte qui suit, tente de répondre à une question posée récemment par Michel Oris et Gilbert Ritschard, (2007), question qui se ramène à celle-ci : dans quelle mesure peut-on affirmer qu'un graphe implicatif présente une « bonne » qualité de structure arborescente par rapport aux données ? Pourrait-on définir une mesure de type inertiel, par exemple, comparable à celle qui permet de qualifier une partition en classes homogènes et « convenablement séparées » ? [1]

Dans la Partie 1, chapitre 4, nous avons établi un critère probabiliste permettant de qualifier des niveaux de la hiérarchie orientée de R-règles, puis celle de la hiérarchie entière. Dans l'ouvrage *L'implication statistique. Nouvelle méthode exploratoire de données* (Gras et al, 1996), nous avons défini un critère numérique, « La variance statistique de classes cohésitives », pour quantifier la qualité d'une hiérarchie de R-règles à ses différents niveaux

[1] Je cite : « …limites de l'ASI….absence de critère permettant de juger de la pertinence statistique globale du modèle retenu. Quel pourrait être un équivalent de la déviance utilisée en modélisation statistique ou de la part d'inertie reproduite en analyse factorielle ? » (M.Oris et G. Ritschard, dans « Dynamique professionnelle dans la Genève du 19ème, enseignements d'une analyse de statistique implicative", Actes de ASI 4,, octobre 2007)

de construction ascendante. Nous procéderons de façon comparable pour quantifier l'état d'homogénéité d'un graphe implicatif[2] selon le seuil qui permet de l'établir.

## 2 Une remarque dans le cas où les variables sont binaires

Considérons un tableau de séries statistiques portant sur un ensemble de n sujets présentés de telle façon que soient regroupés respectivement ceux sur lesquels a et b sont vérifiées simultanément, puis ne le sont pas, puis successivement la variable a l'est sans que ne b le soit, puis la variable b sans que a ne le soit.

Dans le tableau ci -dessous, on donne un exemple d'une telle réorganisation pour n=9.

| Sujets | a | b | Sujets | a | b |
|---|---|---|---|---|---|
| i1 | 1 | 1 | i6 | 1 | 0 |
| i2 | 1 | 1 | i7 | 0 | 1 |
| i3 | 0 | 0 | i8 | 0 | 1 |
| i4 | 0 | 0 | i9 | 0 | 1 |
| i5 | 0 | 0 | | | |

TAB.1

Aux variables a et b, associons les vecteurs-colonnes de leurs occurrences dans l'espace $[0 ; 1]^9$. Le vecteur $\overrightarrow{ab}$ qui se déduit des données de ce couple de variables a pour carré de longueur la valeur 4. Ce nombre caractérise d'évidence la différence symétrique (le « ou » exclusif) awb de a et b (c'est-à-dire la proposition « soit a, soit b »). Mais à ce titre, il est un indicateur de la façon dont a et b s'opposent ou ne s'impliquent pas de l'un vers l'autre, comme $a \wedge \overline{b}$ (resp. $b \wedge \overline{a}$) est un indicateur de a=>b (resp. b=>a). Plus généralement, le carré scalaire du vecteur associé à un couple de variables caractérise la non-implication de l'une sur l'autre. C'est une fonction croissante de l'opposition, de la répulsion d'une variable envers l'autre. Mais, corrélativement, elle décroît avec leur consistance[3], c'est-à-dire une certaine ressemblance. Par exemple, si le carré scalaire est nul, cela signifie que les deux variables sont identiques : les n sujets ont un comportement absolument semblable vis-à-vis de celles-ci. Nous utiliserons ces remarques dans ce qui suit pour donner du sens à la notion de « variance implicative ».

Ces remarques peuvent également s'appliquer au cas où les variables ne seraient plus binaires. Certes la référence à la différence symétrique tombe. Cependant, le carré scalaire du vecteur associé à deux variables numériques continue à exprimer qualitativement et quantitativement l'opposition ou la ressemblance entre les deux variables : plus les instances diffèrent, plus grand est le carré scalaire. Il y aura donc encore correspondance croissante entre « opposition entre deux variables » et « carré scalaire des vecteurs associés ».

[2] Rappelons qu'un graphe implicatif est orienté, sans cycle, pondéré par les intensités d'implication.

[3] Nous employons le mot "consistance" pour éviter l'usage courant, connoté, de "similarité", expressément défini en Analyse de Données, même si les sens en l'occurrence sont voisins

# 3 Un exemple traité par CHIC[4]

Voici un tableau récapitulant les données relatives à 5 variables binaires a, b, c, d et e, observées sur un ensemble de 20 sujets de $i_1$ à $i_{20}$.

| Sujets | **a** | **b** | **c** | **d** | **e** | Sujets | **a** | **b** | **c** | **d** | **e** | Sujets | **a** | **b** | **c** | **d** | **e** | Sujets | **a** | **b** | **c** | **d** | **e** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **I01** | 0 | 1 | 1 | 0 | 0 | **I06** | 0 | 1 | 0 | 0 | 0 | **I11** | 0 | 1 | 1 | 0 | 0 | **I16** | 0 | 1 | 0 | 0 | 0 |
| **I02** | 1 | 0 | 0 | 1 | 1 | **I07** | 1 | 0 | 0 | 0 | 1 | **I12** | 1 | 0 | 0 | 1 | 1 | **I17** | 1 | 0 | 0 | 0 | 1 |
| **I03** | 0 | 0 | 0 | 1 | 1 | **I08** | 0 | 1 | 1 | 1 | 0 | **I13** | 0 | 0 | 1 | 1 | 1 | **I18** | 0 | 1 | 1 | 1 | 0 |
| **I04** | 1 | 0 | 1 | 1 | 0 | **I09** | 1 | 0 | 1 | 1 | 1 | **I14** | 1 | 0 | 1 | 1 | 0 | **I19** | 1 | 0 | 1 | 1 | 1 |
| **I05** | 1 | 0 | 0 | 1 | 1 | **I10** | 0 | 0 | 0 | 1 | 1 | **I15** | 1 | 0 | 0 | 1 | 1 | **I20** | 0 | 0 | 0 | 1 | 1 |

| | a | b | c | d | e |
|---|---|---|---|---|---|
| **Occurrences** | 10 | 6 | 9 | 14 | 12 |

TAB. 2 – *Données binaires de 5 variables*

A chaque variable, point affine de V, nous pouvons associer le vecteur-colonne de ses coordonnées. Par suite, un vecteur tel que le vecteur $\overrightarrow{ab}$, est une image de la façon dont les deux variables a et b se contredisent. La valeur de son carré scalaire euclidien est 16. En effet il ressort qu'en analysant les coordonnées du vecteur $\overrightarrow{ab}$ celles-ci font apparaître une plus grande opposition entre les deux variables a et b que l'analyse des coordonnées du vecteur $\overrightarrow{ae}$ ne le montre entre les deux variables a et e.

Ainsi $\overrightarrow{ab} = (1;-1;0;-1;-1;1;-1;1;-1;0;1;-1;0;-1;-1;1;-1;1;-1;0)$ conduit à $\left\|\overrightarrow{ab}\right\|^2 = 16$ tandis que $\overrightarrow{ae} = (0;0;1;1;0;0;0;0;0;1;0;0;1;1;0;0;0;0;0;1)$ conduit à $\left\|\overrightarrow{ae}\right\|^2 = 6$. Cette propriété d'opposition est symétrique, ce que n'est pas l'implication.

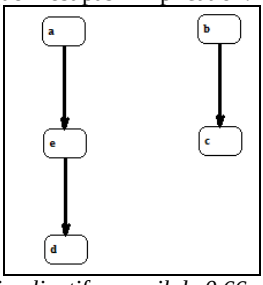


FIG. 1- *Graphe implicatif au seuil de 0,66 relatif à l'exemple*

On obtient les indices d'implication suivants :

[4] CHIC est un logiciel d'analyse de données, développé par Raphaël Couturier (Couturier, 2005) et qui sera présenté dans le chapitre 11 de la Partie 2

Indices d'implications : (selon la théorie classique)
Calcul avec la loi binomiale

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | 0 | 5 | 30 | 60 | 79 |
| b | 2 | 0 | 66 | 3 | 0 |
| c | 28 | 64 | 0 | 52 | 5 |
| d | 58 | 11 | 53 | 0 | 70 |
| e | 76 | 3 | 9 | 73 | 0 |

TAB. 3 - *Intensités d'implication données par CHIC*

Si les variables a et e s'opposent de la même façon, par contre, a implique plus e que la réciproque.

# 4 Formalisation

Nous procéderons comme pour l'implication de classes de variables (Gras et al., 1996) en prenant en compte, à l'instar des méthodes de « clustering », la relation entretenue à travers deux cas entre les éléments d'une même classe (relation intra), leur barycentre et les barycentres respectifs des classes et le barycentre de l'ensemble de ces classes (relation inter). Ainsi, quand on examine des graphes implicatifs restitués par CHIC, on constate que ceux qui sont obtenus sont de deux sortes : ou bien ils se présentent d'un seul « tenant », les chemins du graphe formant un ensemble connexe ou bien, ils sont constitués de sous-ensembles connexes mais mutuellement disjoints. Par exemple, la figure 2 présente deux chemins disjoints (a→e→d) et (b→c) formant un graphe d'un seul « tenant ».

**Définition 1** :
On appelle **chemin implicatif** toute suite de variables, connexe ordonnée par les occurrences croissantes des variables d'origine un nœud sans antécédent et d'extrémité un nœud sans successeur.
**Définition 2 :**
On appelle **grappe implicative** un ensemble connexe de chemins implicatifs connexes.

Au sein de la grappe, deux chemins peuvent partager les mêmes nœuds et les mêmes arcs. Ce partage signifie généralement à la fois une certaine ressemblance sémantique de ces chemins, mais aussi des nuances suffisantes pour que leur coexistence ait un sens dans la grappe. Par exemple, voici un graphe relatif à 6 variables constitué d'une seule grappe :

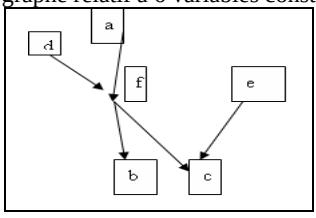


FIG. 2

(d→f→c), (d→f→b), (a→f→b), (a→f→c) et (e→c) forment, en une grappe unique, l'ensemble des 5 chemins du graphe ci-dessus. La figure 1 est, en revanche, constituée de deux grappes, réduites chacune à un chemin.

Notre objectif est de construire une expression qui, associée à un graphe présentant une ou plusieurs grappes, permette d'associer une mesure à la qualité de la structure graphique et, par là, l'homogénéité ou l'hétérogénéité de cette structure en termes de concepts implicitement sous-jacents aux relations entretenues par les variables.

Considérons maintenant un ensemble de n sujets sur lequel nous étudions des caractères représentés par des variables binaires, par l'Analyse Statistique Implicative.

## 4.1 1er cas : le graphe est constitué d'une grappe

A chaque arc du graphe, tel que (d→f), est associé le vecteur $\overrightarrow{df}$ de $R^n$ des composantes des nœuds de ses extrémités d et f. Comme nous l'avons dit, la longueur de ce vecteur caractérise l'intensité de répulsion mutuelle des variables, c'est-à-dire de la non-implication de l'une sur l'autre. Plus précisément, la longueur est fonction décroissante de l'intensité implicative mesurée par le nombre de contre-exemples à l'implication.

Soit p le nombre de chemins composant une grappe.

**Définition 3 :**

On appelle **puissance implicative** du chemin ch(i), la quantité $\varphi_i$ qui est la moyenne géométrique[5] des intensités d'implication des arcs composant le chemin ch(i), y compris les arcs obtenus par transitivité.

Cette puissance implicative a pour effet d'homogénéiser l'ensemble des barycentres des chemins du graphe puisqu'elle diminue leur variance interclasse traditionnelle.

**Définition 4 :**

On appelle **nodulosité** du chemin ch(i), le nombre $k_i$ de variables ou nœuds y figurant.

On ne fait pas ici référence à la notion de longueur du chemin afin de pour pouvoir étendre la notion de chemin à celui réduit à une seule variable. Par convention, une variable isolée, non encore introduite dans le graphe implicatif, peut être considérée comme un chemin de nodulosité 1. Par exemple, le chemin (d→f→c) est de nodulosité 3.

Soit $g_i$ le barycentre du chemin ch(i), c'est-à-dire le barycentre des $k_i$ nœuds (ou variables) du chemin. Le barycentre G de l'ensemble des chemins est le barycentre des p chemins ch(i), pondérés par leur nodulosité $k_i$, constituant le graphe à un seuil choisi 1-$\alpha$.

Considérons la forme bilinéaire symétrique $\Phi$ définie par la matrice carrée diagonale D d'élément générique $\varphi_i$ et de dimension p. La forme quadratique associée définit à son tour une distance $\| \ \|_\Phi$ sur l'ensemble des vecteurs associés aux variables :

$$\Phi(\overrightarrow{g_iG}) = {}^t(\overrightarrow{g_iG})D(\overrightarrow{g_iG}) = \left\|\overrightarrow{g_iG}\right\|_\Phi^2 = \varphi_i^2 . \left\|\overrightarrow{g_iG}\right\|^2 \text{ où } \| \ \| \text{ est la norme euclidienne}$$

[5] Nous choisissons la moyenne géométrique de préférence à la moyenne arithmétique car la première peut alerter d'une valeur faible ou nulle et sera donc plus sensible aux variations des intensités.

Comme nous voulons, par la variance implicative, rendre compte de la façon dont les chemins de variables sont sémantiquement ou contextuellement distingués entre eux ou liés les uns aux autres, nous pondérons, à travers Φ, le carré scalaire ordinaire (euclidien) du vecteur en jeu par la puissance implicative : la différence entre un chemin et l'ensemble des autres est d'autant plus grande que la structure implicative de ce chemin est significativement et relativement importante. Une différence sémantique peut donc être mise en valeur entre deux chemins partageant en commun quelques nœuds et les arcs les reliant.

**Définition 5 :**

On appelle **variance intra-grappe** ou **variance implicative d'une grappe d'un graphe implicatif** admettant p chemins ch(i), la quantité $I_a = \frac{1}{\sum_{i=1}^{i=p} k_i \|\vec{u}_i\|_\Phi^2} \sum_{i=1}^{i=p} k_i \left\|\overrightarrow{g_i G}\right\|_\Phi^2$ où $\|\vec{u}_i\|_\Phi^2 = \varphi_i^2$ est interprétable comme carré de l'intensité d'implication d'un chemin dont toutes les composantes coïncideraient avec celles du barycentre commun G sauf la i$^{ème}$ qui en diffèrerait de 1 et où $k_i \varphi_i^2$ a la signification d'une pondération implicative du chemin

En définitive, nous retrouvons la définition d'une variance statistique intra-classe ordinaire si l'on considère que les quantités $\frac{k_i \varphi_i^2}{\sum_{i=1}^{i=p} k_i \|\vec{u}_i\|_\Phi^2}$ sont des coefficients de somme 1, pondérant les chemins par leur qualité implicative. Mais, pour éviter les confusions et rappeler leur adéquation au problème, nous avons choisi la dénomination **variance intra-grappe**. Et l'expression $I_a$ est un indicateur de dispersion des barycentres des chemins autour du barycentre commun de la grappe. Elle est minimale et nulle lorsque tous les barycentres coïncident avec le barycentre commun. Ce qui est le cas, si la grappe est constituée d'un seul chemin, c'est-à-dire lorsque toutes les variables sont alignées le long du chemin unique.

Les variables isolées, constituant un chemin de nodulosité 1 et une grappe réduite à un chemin auront manifestement une variance intra-grappe nulle. La collection des chemins est d'autant plus porteuse de sens divers et forts que ceux-ci s'opposent dans l'ensemble du graphe. Ce phénomène est accompagné d'une variance intra-grappe importante. En revanche, que deux de ces chemins, ou un réseau de chemins ne s'opposent pas globalement peut souligner l'existence d'une signification commune que le chercheur se devra d'identifier. Ce sera le cas où cette variance sera faible. C'est notamment le cas où la grappe se ramène à un chemin unique.

## 4.2 2$^{ème}$ cas : le graphe est constitué de m grappes

1. Soit $G_1$, $G_2$,…, $G_m$ les barycentres respectifs des m grappes. Le barycentre $G_j$ de la grappe j est affecté de la somme des pondérations des barycentres de ses chemins ;
2. Soit $p_1$,…, $p_m$ les nombres respectifs de chemins implicatifs constituant chacune de ces grappes,
3. Soit $\Gamma$ le barycentre commun des m grappes,

4. Soit $\psi_j$ la moyenne géométrique des puissances implicatives des différents chemins de la grappe j. Cette moyenne $\psi_j$ contient donc une information résumée et relative aux différentes intensités associées aux arcs de la grappe.

Rappelons qu'une variable isolée est un chemin de nodulosité 1 et donc aussi une grappe constituée d'un seul chemin. La puissance implicative d'un tel chemin se réduit à l'intensité affectée à la variable en l'occurrence ici, par définition, elle est égale à 1.

On note $\psi$ la forme bilinéaire symétrique définie par la matrice carrée diagonale $\Delta$ d'élément générique $\psi_j$ et de dimension m. La forme quadratique associée définit à son tour une distance $\| \ \|_\Psi$ sur l'ensemble des vecteurs associés aux vecteurs $\overrightarrow{G_j\Gamma}$.

**Définition 6 :**

On appelle **variance implicative inter-grappe** ou **variance expliquée d'un graphe implicatif** constitué de m grappes, la quantité $I_e = \dfrac{1}{\sum_{j=1}^{j=m} p_j \left\|\vec{v}_j\right\|_\Psi^2} \sum_{j=1}^{j=m} p_j \left\|\overrightarrow{G_j\Gamma}\right\|_\Psi^2$ où $\left\|\vec{\upsilon}_j\right\|_\Psi^2 = \psi_j^2$

est interprétable comme carré de l'intensité d'implication d'une classe dont toutes les composantes coïncideraient avec celles du barycentre commun $\Gamma$ sauf la j$^{ème}$ qui en diffèrerait de 1.

**Définition 7 :**

On appelle **variance implicative totale** du graphe la somme des variances intra-grappes et de la variance inter-grappe

Une variable isolée x a un apport égal à $\left\|\overrightarrow{x\Gamma}\right\|^2$, c'est-à-dire le carré de la distance euclidienne de x à $\Gamma$.

Selon une remarque déjà exprimée, $I_e$ souligne les répulsions mutuelles entre les grappes et donc assure bien la fonction traditionnelle de la variance inter-classe dont il partage la dimension et la forme. En choisissant la dénomination **inter-grappe**, nous voulons cependant éviter les confusions. La somme de la variance intra-grappe et de la variance inter-grappe contient l'information globale implicative restituée par le graphe. Ce qui nous autorise l'appellation choisie de **variance implicative totale**. Les grappes constituées d'un grand nombre de chemins l'affectent a priori plus que les grappes à chemin unique. Si l'homogénéité autour du barycentre global est importante, cela signifie que les grappes ont des sens sans doute voisins et sont potentiellement en passe de se réunir sous la contrainte d'un seuil de confiance de niveau plus faible. Ceci justifie encore a posteriori l'appellation choisie pour la variance implicative.

Ici, comme avec certaines méthodes de classification habituelles, nous apprécions l'homogénéité de l'ensemble des classes par une recherche de minimisation de la variance implicative intra-grappe. Ce qui signifie encore, a contrario, que l'hétérogénéité sera d'autant appréciée que la variance totale sera importante, voire maximale. Ce qui peut être interprété comme signifiant l'existence de concepts différents, voire opposés, au sein de la structure exprimée en graphe. Ce qui, a posteriori encore, justifie d'autant plus la recherche d'une structure significative encore inapparente de l'ensemble des variables.

Supposons qu'à un seuil de construction d'un graphe apparaissent p grappes. Abaissons ce seuil. Apparaissent alors de nouveaux arcs d'intensités nécessairement inférieures

conformément à l'algorithme de construction. Ces arcs peuvent s'agréger aux chemins existants ou constituer de nouveaux arcs tout en supprimant éventuellement l'isolement de variables non encore reliées. La dispersion des centres diminuant, la faiblesse des intensités de ces variables isolées s'amortissant à travers les moyennes géométriques $\varphi_i$, en toute hypothèse, la variance implicative diminuera. En tant que correspondant à une variance inter-classe, elle décroît donc lorsque de nouvelles grappes se forment.

# 5 Retour sur l'exemple numérique

Dans l'exemple donné dans le TAB.2, et représenté par la Fig. 1, nous observons sur le graphe au seuil donné (1-a=0,66) m=2 grappes chacune constituée d'un seul chemin. Pour la grappe 1, le chemin ch(1)= (a→e→d) est de nodulosité 3 et pour la grappe 2 le chemin ch(1)= (b→c) est de nodulosité 2.

Le barycentre de l'ensemble des 5 variables vérifie la relation vectorielle :

$$\overrightarrow{O\Gamma} = \frac{1}{5}(\overrightarrow{Oa} + \overrightarrow{Ob} + \overrightarrow{Oc} + \overrightarrow{Od} + \overrightarrow{Oe})$$

et a pour coordonnées dans $R^{20}$ $\overrightarrow{O\Gamma} = \frac{1}{5}(2;3;2;3;3;1;2;3;4;2;2;3;3;3;3;1;2;3;4;2)$

Par un raisonnement analogue, nous obtenons les barycentres respectifs des deux grappes, $G_1$ et $G_2$

$\overrightarrow{OG_1} = \frac{1}{3}(0; 3; 2; 2; 3; 0; 2; 1; 3; 2; 0; 3; 2; 2; 3; 0; 2; 1; 3; 2)$

$\overrightarrow{OG_2} = \frac{1}{2}(2; 0; 0; 1; 0; 1; 0; 2; 1; 0; 2; 0; 1; 1; 0; 1; 0; 2; 1; 0)$

La puissance implicative du chemin ch(1) de la grappe 1 est : $\varphi_1=(0,79 \text{ x } 0,73 \text{ x } 0,60)^{1/3}$ = 0,702. Comme la grappe 1 ne contient qu'un chemin, la moyenne géométrique $\psi_1$ des puissances des chemins de la grappe 1 coïncide avec la valeur $\varphi_1$. Le carré de la norme euclidienne du vecteur $\overrightarrow{G_1\Gamma}$ est $\left\|\overrightarrow{G_1\Gamma}\right\|^2 = 1,631$

La puissance implicative du chemin ch(1) de la grappe 2 est : $\varphi_{1=}0,66$. Et $\psi_2$ de la grappe 2 coïncide avec $\varphi_{1=}0,66$. Le carré de la norme euclidienne du vecteur $\overrightarrow{G_2\Gamma}$ est $\left\|\overrightarrow{G_2\Gamma}\right\|^2 = 3,67$

Les variances intra-grappes des deux grappes sont nulles puisque leurs barycentres coïncident avec les barycentres des chemins qui les composent. Ainsi, puisque $p_1=p_2=1$, la contribution à la variance inter-grappe de $G_1$ est

$$p_1\left\|\overrightarrow{G_1\Gamma}\right\|_\Psi^2 = p_1\Psi_1^2\left\|\overrightarrow{G_1\Gamma}\right\|^2 = (0,702)^2(1,631) = 0,8039$$

et celle de $G_2$ est $p_2\left\|\overrightarrow{G_2\Gamma}\right\|_\Psi^2 = p_2\Psi_2^2\left\|\overrightarrow{G_2\Gamma}\right\|^2 = (0,66)^2(3,67) = 1,598$

Ces deux nombres expriment l'intensité de répulsion respective de $G_1$ et de $G_2$ sur $\Gamma$.

$$I_e = \frac{1}{\sum_{j=1}^{j=m} p_j \left\| \vec{v}_j \right\|_\Psi^2} \sum_{j=1}^{j=m} p_j \left\| \overrightarrow{G_j \Gamma} \right\|_\Psi^2 = \frac{1}{(0{,}702)^2 + (0{,}66)^2} (0{,}8039 + 1{,}598) = 2{,}58$$

Ce nombre exprime d'une certaine façon l'opposition sémantique entre les deux grappes pour le niveau de confiance de 0,66.

La grappe 1 et la grappe 2 expliquent respectivement 33,5% et 66,5% de la variance implicative totale comme le montre le calcul ci-dessous :.

$$I_T = I_a + I_e = 0 + I_e = \frac{0{,}8039}{(0{,}702)^2 + (0{,}66)^2} + \frac{1{,}598}{(0{,}702)^2 + (0{,}66)^2} = 2{,}58$$

$$\frac{\frac{0{,}8039}{(0{,}702)^2 + (0{,}66)^2}}{2{,}58} = 0{,}3346 \quad \text{et} \quad \frac{\frac{1{,}598}{(0{,}702)^2 + (0{,}66)^2}}{2{,}58} = 0{,}6654$$

Pour poursuivre notre illustration, nous abaissons le seuil de confiance au niveau à 0,52[6], et conséquemment apparaît un nouveau graphe implication réduit à une seule grappe mais comportant deux chemins, c'est à dire que m=1 et p=2. Le chemin ch(1) de nodulosité $k_1$=4 et le chemin ch(2), $k_2$=2.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | 0 | 0,05 | 0,30 | **0,60** | **0,79** |
| b | 0,02 | 0 | **0,66** | 0,03 | 0 |
| c | 0,28 | 0,64 | 0 | 0,52 | 0,05 |
| d | 0,58 | 0,11 | **0,53** | 0 | 0,70 |
| e | 0,76 | 0,03 | 0,09 | **0,73** | 0 |

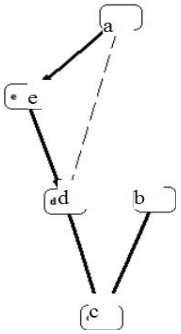


FIG. 3 - *Graphe implicatif au seuil de 0,52 relatif à l'exemple du tableau 2*

La puissance implicative du chemin ch(1)=(a, e, d, c) est de :

$\varphi_1 = (0{,}79\text{x}0{,}73\text{x}0{,}60\text{x}0{,}53)^{\frac{1}{5}} = 0{,}7123$. Celle du chemin ch(2)=(b, c) est $\varphi_2 = 0{,}66$

et les coordonnées de son barycentre $g_2$ sont

$$\overrightarrow{Og_2} = (1, 0, 0, 1/2, 0, 1/2, 0, 1, 1/2, 0, 1, 0, 1/2, 1/2, 0, 1/2, 0, 1, 1/2, 0)$$

Le barycentre commun reste le même que dans la situation précédente et comme il n'y a qu'une grappe, son barycentre G coïncide avec le barycentre global Γ :

$$\overrightarrow{OG} = \overrightarrow{O\Gamma} = \frac{1}{5}(2;3;2;3;3;1;2;3;4;2;2;3;3;3;3;1;2;3;4;2) .$$

[6] Le seuil est très bas en raison du faible nombre d'occurrences. Mais il s'agit d'un simple exemple d'école.

Les coordonnées du barycentre $g_1$ de (a, e, d, c) sont :

$$\overrightarrow{Og_1} = \frac{1}{4}(1, 3, 2, 3, 3, 0, 2, 2, 4, 2, 1, 3, 3, 3, 3, 0, 2, 2, 4, 2)$$

$$I_T = I_a + I_e = I_a + 0 = \frac{1}{k_1\varphi_1^2 + k_2\varphi_2^2}\left[ k_1\varphi_1^2 \left\|\overrightarrow{g_1G}\right\|^2 + k_2\varphi_2^2 \left\|\overrightarrow{g_2G}\right\|^2 \right]$$

$$I_T = I_a = \frac{1}{4(0{,}7123)^2 + 2(0{,}66)^2}\left[ 4(0{,}7123)^2 \left\|\overrightarrow{g_1G}\right\|^2 + 2(0{,}66)^2 \left\|\overrightarrow{g_2G}\right\|^2 \right]$$

avec $\left\|\overrightarrow{g_1G}\right\|^2 = 0{,}4325$ et $\left\|\overrightarrow{g_2G}\right\|^2 = 3{,}67$

Par conséquent $I_T = I_a = 1{,}4048$ . La contribution relative du chemin ch(1) à la variance intra-grappe est de 21,5% et celle du chemin ch(2) est de 78,5%

Comme la variance implicative inter-grappe est nulle, la variance implicative totale est donc 1,40 c'est-à-dire plus faible qu'au seuil précédent Au niveau de faible exigence relationnelle (0,52) l'hétérogénéité entre les deux chemins est atténuée grâce à la connexion établie entre les deux chemins précédemment indépendants. Mais l'information globale, restituée par la variance, s'est appauvrie car il est possible que les sens distincts de ces deux chemins se soient quelque peu dilués en un sens moins spécifique.

# 6 Autre exemple[7]

Sur un ensemble de 270 hommes et femmes, on dispose d'informations portant sur leur état-civil, et leur secteur d'activité. La problématique des chercheurs est la recherche de critère de décision à partir de prédicteurs à l'aide de la méthode d'arbre de décision.

| Etat civil | Homme | | | Femme | | | Total |
|---|---|---|---|---|---|---|---|
| | primaire | secondaire | tertiaire | primaire | secondaire | tertiaire | |
| Marié | 50 | 40 | 6 | 0 | 14 | 10 | 120 |
| Célibataire | 5 | 5 | 12 | 50 | 30 | 18 | 120 |
| Divorcé/veuf | 5 | 8 | 10 | 6 | 2 | 2 | 33 |
| Total | 60 | 53 | 28 | 56 | 46 | 30 | 270 |

TAB. 4 -

La variable à prédire est l'état civil, le sexe et le secteur d'activité étant les prédicteurs disponibles. Une analyse statistique implicative appliquée à cette question conduit aux implications suivantes :

[7] Cet exemple est extrait de l'article de G.Ritschard, D.Zighed et S.Marcellin [2007], Données déséquilibrées, entropie décentrée et indice d'implication, Nouveaux apports théoriques à l'Analyse Statistique Implicative et Applications 4èmes Rencontres Internationales d'Analyse Statistique Implicative, Université Jaume I de Castellon 18-21/10/2007, p. 315-328, ISBN 978-84-690-8241-6

| règles | Intensités d'implication | |
|---|---|---|
| **S**econdaire => **M**arié | φ (**S**,**M**) = 0,94 | |
| **H**omme => Marié | φ (**H, M**) = 0,99 | $\varphi_{entropique}$ (**H, M**) =0,68 |
| **T**ertiaire => **C**élibataire | φ (**T,C**) = 0,78 | |
| **F**emme => **C**élibataire | φ (**F,C**) = 0,99 | $\varphi_{entropique}$(**F,C**) =0,78 |

TAB. 5

Autrement dit, on peut ne pas prendre de grand risque d'affirmer que si un sujet de l'enquête est une femme (resp. un homme) elle (resp. il) est célibataire (resp. marié). Si l'échantillon interrogé respecte les règles de tirage au hasard dans une population plus large, l'induction de ces propriétés est légitime.

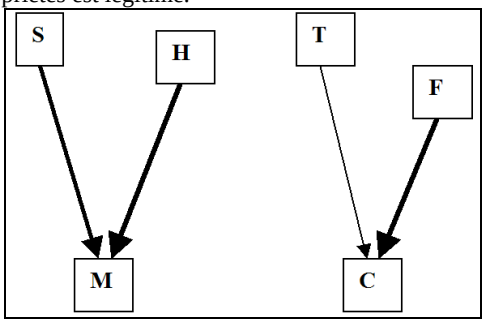


FIG. 4 - *Graphe implicatif des 4 variables au seuil 0,778*

Il s'agit ici de deux grappes composées chacune de deux chemins. Chaque chemin est de nodulosité égale à 2. Les calculs conduisent aux résultats suivants :

* la grappe 1 (**S,H,M**) a une variance intra-grappe égale à 8,954
* la grappe 2 (**T,F,C**) a une variance intra-grappe égale à 8,125
* ainsi, la somme des variances intra-grappes $I_a$ est de 17,08
* la variance inter-grappe $I_e$ est la somme de la contribution de la grappe 1 (**S,H,M**) soit $\frac{62}{2(0,93^2+0,77^2)} = 21,26$ et de celle de la grappe 2 (**T,F,C**) soit $\frac{51,44}{2(0,93^2+0,77^2)} = 17,64$.

D'où $I_e$ = 38,91.

La grappe 1 **(S,H,M)** explique 54,64% de la variance inter-grappe, soit plus que la grappe (**T,F,C**) dont la structure est quelque peu affaiblie par l'intensité implicative φ (**T,C**).

On note alors que la variance *inter-grappe* est plus forte que la variance *intra-grappe* ce qui souligne la bonne cohérence interne de chacune des grappes mais leur opposition sémantique, évidemment perceptible.

**Remarque**

A titre indicatif, nous avons pratiqué une analyse statistique implicative à partir de certains itemsets, en l'occurrence, ceux qui associent le sexe et le secteur d'activité. On observe alors que la faible liaison entre P et C (intensité d'implication : 0,69) est très sensiblement améliorée en conjoignant P et F : (P, F) => C avec une intensité de 1. Ceci signifie qu'une Femme ayant des activités dans le secteur Primaire est presque sûrement

Célibataire. Une analyse en termes de variance implicative pourrait être conduite sur le nouveau graphe qui serait composé, cette fois, de deux grappes de trois chemins.

# 7 Conclusion : rôle descriptif et rôle décisionnel de la variance implicative

Nous avons vu, grâce à un exemple, que, pour un seuil 1-$\alpha$ du graphe implicatif donné, la variance implicative est une mesure qu'il est possible d'associer au graphe. La valeur de la fraction intra peut être comparée à la variance implicative totale et donner ainsi une information quantitative sur la structure du graphe au seuil retenu. Le rapport entre la variance implicative intra-grappe au seuil 1-$\alpha$ et la variance implicative totale en rend compte de façon claire mais non décisive. On peut en observer les variations en abaissant le seuil et exploiter les « vitesses » respectives de décroissance entre le seuil et la variance.

Notons aussi que la variance implicative, contrairement à l'intensité d'implication qui est une probabilité, ne fournit pas d'échelle de mesure. Par exemple, l'appréciation d'un éventuel antagonisme entre grappes ne peut s'estimer que par comparaison relative. En effet, toujours à un seuil donné, les « contributions » respectives des chemins à la variance apportent une information qui permet, par comparaison, de décider de la qualité contributive de la variance associée au chemin à la variance implicative totale. Et donc de pointer le chemin le plus structuré. L'attribution d'un sens à ce chemin peut donc être plus assurée qu'elle ne peut l'être pour les autres.

Il est possible également d'associer seuil et variance. Comme plus le seuil est abaissé, plus le nombre de chemins s'accroît de façon naturelle, il serait intéressant d'adopter un critère de significativité de la structure en formant le produit variance et seuil. Ainsi la croissance de la variance serait compensée, dans sa signification, par celle du seuil. Relativement à l'exemple donné dans le sous-chapitre 3 et traité dans le 5, nous aurions donc à comparer les variances respectives 2,58 au seuil 0.66 et 1,40 au seuil 0,52. Les deux produits respectifs 1,702 et 0,728 confirment que la deuxième structure présente une valeur inférieure et que le regroupement en deux chemins contigus est chargé de moins de sens que le graphe en deux chemins disjoints, comme si ce sens s'était affadi, avait perdu ses nuances discriminatoires.

Sur le plan décisionnel, il ne paraît pas possible, compte tenu des paramètres en jeu comme l'intensité d'implication, le nombre de chemins, le seuil de construction du graphe, de définir quelle est la structure la plus intéressante parmi celles qui peuvent être obtenues : faut-il minorer la variance implicative intra-grappe afin de disposer d'une bonne qualité de leur architecture qui renforcerait leur sens ? ou faut-il majorer la variance implicative inter-grappe afin de s'assurer d'une forte opposition des significations des grappes ? En fait, interactivement, ce sont ces deux optimisations qu'il serait bon de préserver en laissant l'expert apprécier leurs variations au fil de celle du seuil. La décision d'arrêt dépendra de la qualité de l'expression rendue par la structure graphique intégrant le doute lié à un seuil assez bas et la certitude d'une discrimination sur un sens suffisamment fort.

L'extension de cette approche quantitative pour évaluer la qualité structurelle d'un graphe au cas où les variables ne seraient pas binaires nous paraît légitimée par la référence à la différence symétrique, métaphore de la répulsion ou de la consistance.

Sur le plan formel, une autre approche semble possible et d'intérêt manifeste : Gilbert Ritschard suggère, de son côté, une approche à l'aide de l'entropie décentrée. De nature

foncièrement différente, la formalisation conduirait à des résultats qu'il serait intéressant à comparer aux précédents.

## Références

Couturier R. et R. Gras (2005) : CHIC : Traitement de données avec l'analyse implicative, *Extraction et Gestion des Connaissances, Volume I1, RNTI,* Cépaduès, Paris, p.679-684, ISBN 2.85428.683.9.

Gras R. (1979) *Contribution à l'étude expérimentale et à l'analyse de certaines acquisitions cognitives et de certains objectifs didactiques en mathématiques*, Thèse d'Etat, Université de Rennes I.

Gras R., S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn et A. Totohasina (1996), *L'implication Statistique*, Grenoble : La Pensée Sauvage.

Oris M. et G. Ritschard (2007) , Dynamique professionnelle dans la Genève du 19ème siècle ; enseignements d'une analyse de statistique implicative, *Nouveaux apports théoriques à l'Analyse Statistique Implicative et Applications*, 4èmes Rencontres Internationales d'Analyse Statistique Implicative, Castellon, 2007, ISBN 978-84-690-8241-6, 287-300.

Ritschard G., D. Zighed et S. Marcellin (2007), Données déséquilibrées, entropie décentrée et indice d'implication, *Nouveaux apports théoriques à l'Analyse Statistique Implicative et Applications* 4èmes Rencontres Internationales d'Analyse Statistique Implicative, Université Jaume I de Castellon 2007, ISBN 978-84-690-8241-6, 315-328.

## Summary

A weighted graph, without cycle, is a representation of a set of implicative association rules extracted from a digital cross variables and subjects. The problem of homogeneity, consistency and therefore the relevance of the interpretations of expert arises since in Statistical Implicative Analysis (SIA) it is possible to vary the threshold of representation of partial rules. We present here the concept of variance implicative like the classical concept of variance in order to characterize the homogeneity of the representation. It relies on a metaphor of repulsion versus mutual consistency between two binary variables from their symmetric difference.