NOTION DE CHAMP IMPLICATIF EN ANALYSE STATISTIQUE IMPLICATIVE

Régis GRAS, Pascale KUNTZ et Nicolas GREFFARD[1]

RÉSUMÉ

Dans le cadre de la théorie de l'Analyse Statistique Implicative, la problématique de la stabilité de l'indice qui permet de définir et évaluer la qualité de l'indice d'implication est posée par l'utilisateur qui renouvelle ses expériences dans un domaine particulier. Dans cet article, nous étudions ce problème en invoquant les concepts différentiels de l'analyse mathématique. Nous examinons un à un les paramètres intervenant dans la formule donnant l'indice d'implication. Nous comparons les variations de ces paramètres avec ceux d'autres indices classiques utilisés en fouille de données. Nous étendons cette étude par celle de la structure de l'espace vectoriel qu'ils engendrent et en centrant cette étude sur la notion de gradient implicatif. De là, nous illustrons par une représentation géométrique la problématique de l'équilibre de l'indice via une discrétisation des surfaces équipotentielles.

Mots-clés : *analyse statistique implicative**,** intensité d'implication, indice d'implication, différentielle, gradient, surface équipotentielle*

ABSTRACT

In the context of the theory of implicative statistical analysis, a user repeating some experimentation in a specific domain is faced with the issue of the robustness of the metric appraising the quality of the implicative index. In this paper, we address this problem through a differential analysis instead of bootstrapping. We study each individual parameter involved in the implicative index equation. And we compare their variations with those of other indices from the data-mining literature. Furthermore, we study the structure of the vector field they span by focusing on the notion of implicative gradient. From there, a geometrical representation is used to illustrate the index equilibrium problematic through a series of figures of equipotential surfaces.

Keywords *: statistical implicative analysis**,** implication intensity, measure of implication, differential, equipotential surface*.

1 Introduction

[1] École Polytechnique de l'Université de Nantes, Équipe DUKE Data User Knowledge, Laboratoire d'Informatique de Nantes-Atlantique (LINA), UMR 6241, Site de la Chantrerie, rue C.Pauc, BP 44306, Nantes cedex 3, e-mail : regisgra@club-internet.fr, pascale.kuntz@univ-nantes.fr et nicolas.greffard@univ-nantes.fr.

Le chercheur en sciences et particulièrement en sciences humaines, expérimentateur ou non, vise, à travers ses interrogations, à construire ou conforter des connaissances dans son domaine. Il est confronté à des situations, vécues par une population de sujets dans lesquelles apparaissent des phénomènes, par exemple des attributs, conduisant à de nombreuses données d'observation. De celles-ci, parmi d'autres informations, il veut extraire des relations d'association entre les variables observées. Puis, à partir de ces relations, il cherche à constituer des savoirs stables et partagés dans son domaine. Parmi celles-ci, les relations causales ou règles non symétriques -« cause => effet (ou conséquence) » - figurent en bonne place pour leur intérêt de découvertes à valeur prédictive. Mais rares sont les réponses respectant les principes métaphysiques de la logique formelle c'est-à-dire s'exprimant en terme de « vrai » ou de « faux ». Ces réponses seraient plus acceptables évaluées dans une logique **dialectique,** « logique des contraires », logique qui permet de dépasser les contradictions. C'est cette composante épistémologique qui guide, de façon originale, la méthode d'analyse de données, l'Analyse Statistique Implicative -d'acronyme ASI-, que nous avons élaborée afin que prévalent les **quasi-règles**, c'est-à-dire celles qui acceptent des contre-exemples sans effacer leur plausibilité. En effet, une règle non strictement satisfaite y trouve sa place en même temps qu'elle serait logiquement réfutable ; autrement dit, simultanément, règle et sa négation ont droit de coexister. La contradiction est provisoirement levée en ASI par la donnée d'un seuil probabiliste variable dont le chercheur contrôle la flexibilité : au-delà, la quasi-règle est acceptée, en-deçà elle est mise à l'écart. Un certain déterminisme probabiliste confère à la quasi-règle une prédictibilité statistique. Autrement dit, comme l'énonce le philosophe Lucien Sève (Sève, 2005, p.104), « on perd en rigueur ce que l'on le gagne en richesse » et, ajouterons-nous, en fécondité.

La méthodologie que nous développons dans l'ASI et que nous implémentons dans le logiciel Classification Hiérarchique Implicative et Cohésitive, d'acronyme CHIC, consiste, en partant de la contingence des données, issues d'un ensemble répétitif de phénomènes, à analyser le singulier, le « surprenant ». Celui-ci qui constitue la quasi-règle, débouche dialectiquement sur une interprétation, une signification et donc une esquisse de savoir général, admis provisoirement comme une réalité indépendante, platonicienne. Car, comme le dit plus loin L. Sève : « il y a toujours de l'universel dans le singulier » (p. 200). Pour ce faire, nous construisons un modèle mathématique qui permet la comparaison statistique du contingent et du théorique aux ordres du modèle. Au sujet de son application possible en biologie, Jean-Claude Ameisen déclare dans « Sur les épaules de Darwin, tome 2» (p. 163) : « *A la recherche des régularités, des mécanismes et des relations de causalité qui rendent compte des formes et des comportements du monde vivant, la formalisation mathématique jouera un rôle de plus en plus important en biologie ».*

Plus précisément, plaçons-nous donc dans le cadre de la théorie de l'Analyse Statistique Implicative, où l'on croise des sujets ou des objets avec des variables de natures variées. Restreignons-nous, pour le moment, au cas de variables binaires telles que a et b. La problématique majeure consiste en la recherche d'une mesure qui permettrait de quantifier la règle quasi-implicative de a sur b. Cette mesure, en ASI, se fonde sur le nombre de contre-exemples à l'implication, nombre qui doit être le plus petit possible eu égard aux cardinaux des sous-ensembles de sujets vérifiant

respectivement a et b (voir « *L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités* » [2013]) . Nous nous intéresserons à l'indice que retient l'ASI pour signifier ou non si la qualité de la règle est statistiquement acceptable et, qu'en conséquence, si la tendance à la règle implicative a de bonnes raisons d'être ou non retenue en étant *étonnamment* acceptable[2] dans le cadre d'une théorie.

Nous examinerons les propriétés de sensibilité de cet indice, dit d'implication, particulièrement aux variations des cardinaux en jeu dans des expériences d'extraction de règles de corpus où les variables sont instanciées sur l'ensemble des sujets. On s'intéressera en particulier au gradient de cet indice significatif de la « vitesse et de l'amplitude » de la croissance ou la décroissance de ces cardinaux. Nous étudierons le champ de vecteurs engendré par les observations de l'indice d'implication et le champ de gradient qui lui est associé. Cette approche diffère donc de celle adoptée par Eugen Barbu dans son mémoire de D.E.A.(Barbu E., 2003) intitulé « Hiérarchie cohésitive » qui, à travers un bootstrap appliqué à des mesures de qualité de règles d'association, a fait varier les paramètres de ces mesures dont celle qui fonde l'implication statistique.

Des modélisations différentes de l'ASI pour la recherche de règles quasi-implicatives se retrouvent sous la forme de treillis de Galois et de réseau Bayesien. Un algorithme dit « A priori » est le plus souvent à la base de la mesure de qualité de règles implicatives. Nous en avons comparé les propriétés face à l'ASI tant sur le plan de leur sensibilité aux variations des instances des variables que sur la représentation des structures des ensembles de règles obtenus. Nous y revenons rapidement dans le § 4. C'est donc plus sur ces phénomènes que nous porterons notre attention et non pas, *dans ce texte,* sur la représentation des règles et méta-règles. Nous rappelons ci-dessous, le processus probabiliste qui en permet l'extraction.

2 Rappel des fondements de l'A.S.I.(Gras, 1979)

Rappelons les fondements premiers de l'ASI dont l'objectif est de rechercher la plausibilité causale derrière une relation quasi-implicative là où le physicien philosophe Bernard d'Espagnat (B.d'Espagnat, 1981) parlerait d**'influence** en ces termes : *« les évènements A influencent appréciablement les évènements B si et seulement si la fréquence avec laquelle les évènements B ont lieu est (appréciablement) différente selon que l'on impose on non aux évènements A d'exister* » (p. 186). Reprenant cette perspective, nous mathématisons cette définition en quantifiant l'expression floue «appréciablement », non pas par une fréquence conditionnelle comme il est fait dans

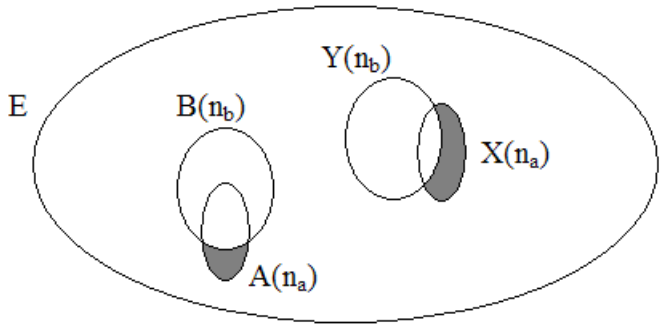
[2] C'est aussi ce qu'affirme René Thom (« Paraboles et catastrophes », 1980, p.130) : « …*le problème n'est pas de décrire la réalité, le problème consiste bien plus à repérer en elle ce qui a de sens pour nous, ce qui est surprenant dans l'ensemble des faits. Si les faits ne nous surprennent pas, ils n'apportent aucun élément nouveau pour la compréhension de l'univers : autant donc les ignorer* » et plus loin : « … *ce qui n'est pas possible si l'on ne dispose pas déjà d'une théorie* ».

(Agrawal R. et al, 1993) et ses dérivés, par exemple dans les réseaux bayesiens, mais de la façon suivante.

Notons A et B les sous-ensembles respectifs de E d'individus qui vérifient respectivement les variables booléennes a et b (Fig. 1). Depuis 1979, nous avons étendu (cf. Ouvrages cités en référence), de façon appropriée, les propriétés qui suivent à des variables réelles continues ou non.

Pour une règle quelconque $a \rightarrow b$, observée dans E, l'ASI consiste à comparer le nombre de contre-exemples $n_{a \wedge \bar{b}}$ ($\bar{b}$ est la négation de b) à cette règle observés dans l'intersection $A \cap \overline{B}$ ($\overline{B}$ est le complémentaire de B dans E) avec le nombre de contre-exemples qui apparaîtraient lors d'un choix aléatoire et indépendant de deux parties X et Y de E de mêmes cardinaux respectifs que A et B (Fig. 1) (Gras, 1979 ; Lebart et al., 2006). La variable aléatoire associée est notée $N_{a \wedge \bar{b}}$. Le principe fondamental que nous retenons consiste à comparer la contingence à la théorie par une méthodologie qui relève de la philosophie de l'expérience où sont premières les observations (cf. B. d'Espagnat, 1981, cité plus haut).

La qualité de la règle sera intuitivement d'autant meilleure que Prob[$N_{a \wedge \bar{b}} > n_{a \wedge \bar{b}}$] sera proche de 1 : autrement dit, dans ce cas, on y observe plus de contre-exemples dans des circonstances aléatoires que l'on en a observés dans la contingence, sous l'hypothèse a priori d'indépendances des variables a et b. Dans ce cas, le seul hasard conduit donc, en moyenne, à plus de contre-exemples que ce qui est observé. C'est ainsi que se définit l'ASI en falsifiant la relation implicative par la mesure relative de sa négation exprimée à travers ses contre-exemples.



Les parties grisées représentent les contre-exemples à l'implication a⇒b

Figure 1. Représentation ensembliste

La méthode de tirage au hasard de X et Y, dans une hypothèse a priori d'indépendance de *a* et *b*, conduit à différentes options pour la loi de la variable aléatoire $N_{a \wedge \bar{b}}$. Deux modélisations de cette variable sont généralement retenues en ASI conduisant à un modèle de Poisson et à un modèle binomial (Gras et al 2009). On centre et on réduit cette variable en la variable $Q(a, \bar{b})$; l'observation contingente (empirique), sa réalisation, est $q(a, \bar{b})$. Par exemple, dans le cas du modèle de Poisson, on obtient l'indice de base pour des variables binaires :

$Q(a,\bar{b}) = \dfrac{Card(X \cap \bar{Y}) - \dfrac{n_a n_{\bar{b}}}{n}}{\sqrt{\dfrac{n_a n_{\bar{b}}}{n}}}$ alors que $q(a,\bar{b}) = \dfrac{n_{a\wedge\bar{b}} - \dfrac{n_a n_{\bar{b}}}{n}}{\sqrt{\dfrac{n_a n_{\bar{b}}}{n}}}$ est sa réalisation contingente

L'intensité d'implication est alors définie par $\varphi(a,b) = Prob[Q(a,\bar{b}) > q(a,\bar{b})]$, dont la valeur gaussienne asymptotique, centrée et réduite est :

$$\varphi(a,b) = \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{+\infty} e^{-\frac{t^2}{2}} dt \quad (1)$$

Cette définition (1) de l'intensité d'implication, rappelle au chercheur qu'elle ne présente un intérêt implicatif qu'à condition qu'elle soit supérieure à 0.50, c'est-à-dire que $q(a,\bar{b})$ soit négatif. Plus $q(a,\bar{b})$ est négatif, meilleure est la qualité de la règle $a \rightarrow b$. Cependant, si $n_b = n$, nous définissons l'intensité d'implication par : $\varphi(a,b) = 0$ qui n'est pas le prolongement par continuité de l'intensité définie par (1) alors que $q(a,\bar{b})$ n'est plus défini pour cette valeur de n_b. Lever son indétermination reviendrait à lui attribuer la valeur 0, et donc $\varphi(a,b) = 0.5$, incompatible avec la sémantique de l'ASI. Cette exception tient au respect de celle-ci puisque l'implication de a sur b est alors devenue triviale, tautologique, donc non informative.

Dans cet article, nous nous intéressons aux propriétés de la fonction q(.) comme fonction des 4 occurrences observées, c'est-à-dire les variables cardinales n, n_a, n_b, $n_{a\wedge\bar{b}}$. Quel rôle jouent ces variations sur celles de l'intensité d'implication qui dépend de q(.) par intégration gaussienne ? Quelle sensibilité observe-t-on lorsque les occurrences varient quelque peu ? En quoi le gradient de q permet-il de définir un champ de gradient sur l'espace des couples de variables ? Nous sommes persuadés que la perception d'un graphe implicatif, passe certes par l'examen de sa structure qui comme l'énergie est de type « intégrale », mais aussi qu'elle se révèle à travers des procédures de dérivation. Sans mésestimer le premier que nous explorerons dans une modélisation originale mécanique de l'ASI, nous consacrerons l'étude qui vient à une approche essentiellement infinitésimale.

3 Variations de l'indice d'implication q en fonction des 4 occurrences

3.1 Stabilité de l'indice d'implication

Étudier la stabilité de l'indice d'implication q, revient à examiner ses petites variations au voisinage des 4 valeurs entières observées (n, n_a, n_b, $n_{a\wedge\bar{b}}$). Pour ce faire, il est possible d'effectuer différentes simulations en croisant ces 4 variables entières dont q dépend (Gras et al., 2013). Mais, considérons ces variables comme variables à valeurs réelles et q comme une fonction continûment différentiable par rapport à ces variables, elles-mêmes contraintes à respecter les inégalités : $0 \leq n_a \leq n_b$ et $n_{a\wedge\bar{b}} \leq \inf\{n_a, n_b\}$ et $\sup\{n_a, n_b\} \leq n$. La fonction q définit alors un champ scalaire et vectoriel sur R^4 en tant qu'espace affine et vectoriel sur lui-même. Dans l'hypothèse vraisemblable d'une évolution d'un processus non chaotique du recueil de données, il suffit alors d'examiner la différentielle de q par rapport à ces variables et d'en conserver la

restriction aux valeurs entières des paramètres de la relation a⇒b. La différentielle de q, au sens de la topologie de Fréchet[3], s'exprime de la façon suivante par un produit scalaire :

$$dq = \frac{\partial q}{\partial n} dn + \frac{\partial q}{\partial n_a} dn_a + \frac{\partial q}{\partial n_b} dn_b + \frac{\partial q}{\partial n_{a\wedge\bar{b}}} dn_{a\wedge\bar{b}} = \text{grad q. dM (2)}^4$$

où M est le point de coordonnées (n, n_a, n_b, $n_{a\wedge\bar{b}}$) du champ scalaire C de vecteurs,

dM est le vecteur de composantes les accroissements différentiels de ces variables d'occurrences,

- et grad q le vecteur de composantes les dérivées partielles de ces variables occurrences.

La différentielle de la fonction q apparaît donc comme le produit scalaire de son gradient et de l'accroissement de q sur la surface représentant les variations de la fonction q(n, n_a, n_b, $n_{a\wedge\bar{b}}$). Ainsi, le gradient de q représente ses propres variations en fonction de celles de ses composantes, les 4 cardinaux des ensembles E, A, B et $A \cap \overline{B}$. Il indique la direction et le sens de croissance ou la décroissance de q dans l'espace de dimension 4. Rappelons qu'il est porté par la normale à la surface de niveau q = cte.

Si l'on veut étudier comment varie q en fonction de $n_{\bar{b}}$, il suffit de remplacer n_b par n-n_b et donc changer le signe de la dérivée de n_b dans la dérivée partielle. En fait, l'intérêt de cette différentielle réside dans l'estimation de l'accroissement (positif ou négatif) de q que nous notons Δq par rapport aux variations respectives Δn, Δn_a, Δn_b, $\Delta n_{\bar{b}}$ et $\Delta n_{a\wedge\bar{b}}$. On a donc :

$$\Delta q = \frac{\partial q}{\partial n} \Delta n + \frac{\partial q}{\partial n_a} \Delta n_a + \frac{\partial q}{\partial n_b} \Delta n_b + \frac{\partial q}{\partial n_{a\wedge\bar{b}}} \Delta n_{a\wedge\bar{b}} + o(\Delta q)$$

où o(Δq) est un infiniment petit du 1er ordre.

Examinons les dérivées partielles de n_b et le nombre de contre-exemples $n_{a\wedge\bar{b}}$. On obtient :

[3] La topologie de Fréchet admet comme base de filtres des sections de N, soit des sous-ensembles de naturels de la forme {n, n+1, n+2, ….} alors que la topologie usuelle sur R admet pour filtres des intervalles de réels. Ainsi continuité et dérivabilité sont des concepts parfaitement définis et opératoires selon la topologie de Fréchet au même titre qu'ils le sont avec la topologie usuelle.

[4] Par une métaphore mécaniste, on dira que dq est le travail élémentaire de q pour un déplacement dM.

$$\frac{\partial q}{\partial n_b} = \frac{1}{2} n_{a\wedge\bar{b}} \left(\frac{n_a}{n}\right)^{-\frac{1}{2}} (n - n_b)^{-\frac{3}{2}} + \frac{1}{2} \left(\frac{n_a}{n}\right)^{\frac{1}{2}} (n - n_b)^{-\frac{1}{2}} > 0 \quad (3)$$

$$\frac{\partial q}{\partial n_{a\wedge\bar{b}}} = \frac{1}{\sqrt{\frac{n_a n_{\bar{b}}}{n}}} = \frac{1}{\sqrt{\frac{n_a (n - n_b)}{n}}} > 0 \quad (4)$$

Ainsi, si les accroissements Δn_b et $\Delta n_{a\wedge\bar{b}}$ sont positifs, l'accroissement de $q(a,\bar{b})$ est également positif. Ceci s'interprète ainsi : si le nombre d'exemples de b et celui des contre-exemples de l'implication augmentent alors l'intensité d'implication diminue pour n et n_a constants. Autrement dit, cette intensité d'implication est maximum aux valeurs observées n_b et $n_{a\wedge\bar{b}}$ et minimum aux valeurs $n_b+\Delta n_b$. et $n_{a\wedge\bar{b}}+\Delta n_{a\wedge\bar{b}}$.

Si nous examinons le cas où n_a varie, nous obtenons la dérivée partielle de q par rapport à n_a qui est :

$$\frac{\partial q}{\partial n_a} = -\frac{1}{2\sqrt{n_{\bar{b}}/n}} \cdot \frac{n_{a\wedge\bar{b}}}{n} \cdot \left(\frac{n}{n_a}\right)^{\frac{3}{2}} - \frac{1}{2}\sqrt{\frac{n_{\bar{b}}}{n_a}} < 0 \quad (5)$$

Ainsi, pour des variations de n_a sur [0, n_b], la fonction indice d'implication $q(a,\bar{b})$ est toujours décroissante (et concave) par rapport à n_a et est donc minimum pour $n_a = n_b$. Par suite, l'intensité d'implication y est croissante et maximum pour $n_a = n_b$.

Notons la dérivée partielle de q par rapport à n :

$$\frac{\partial q}{\partial n} = \frac{1}{2\sqrt{n}} \left[n_{a\wedge\bar{b}} + \frac{n_a n_{\bar{b}}}{n}\right]$$

En conséquence, si les 3 autres paramètres sont constants, l'indice d'implication décroît en $\sqrt{n}$. La qualité de l'implication n'en est donc que meilleure, propriété spécifique de l'ASI par rapport à d'autres indicateurs retenus dans la littérature (cf. Gras et Couturier, 2010). Cette propriété est en accord avec les attentes statistiques et sémantiques relatives au crédit accordé à la fréquence des observations. Les dérivées partielles de q (au moins l'une d'entre elles) étant non linéaires selon les paramètres variables en jeu, on a affaire à un système dynamique non linéaire avec toutes les conséquences épistémologiques que nous envisagerons par ailleurs.

3.2 Exemple numérique

Dans une première expérience, on observe les occurrences :

n = 100, n_a = 20, n_b = 40 (d'où $n_{\bar{b}} = 60$), $n_{a\wedge\bar{b}}$ = 4.

L'application de la formule (1) donne $q(a,\bar{b})$ = - 2,309

Dans une 2ème expérience, n et n_a sont inchangées mais les occurrences des b et des contre-exemples $n_{a\wedge\bar{b}}$ s'accroissent d'une unité.

Au point initial de l'espace des 4 variables, les dérivées partielles qui seules nous intéressent (selon n_b et $n_{a\wedge\bar{b}}$) ont respectivement pour valeurs en appliquant les formules (3) et (4)) : $\frac{\partial q}{\partial n_b} = 0{,}0385$ et $\frac{\partial q}{\partial n_{a\wedge\bar{b}}} = 0{,}2887$

Comme Δn_b, $\Delta n_{\bar{b}}$ et $\Delta n_{a\wedge\bar{b}}$ sont égaux à 1, -1 et 1, l, alors Δq est égal à :

0,0385 + 0.2887 + o(Δq) = 0,3272 + o(Δq) et la valeur approchée de q lors de la deuxième expérience est – 2,309 + 0,2887 + o(Δq)= -1,982 + o(Δq) en utilisant le développement de q au premier ordre (formule (2)).

Or le calcul du nouvel indice d'implication q au point de la 2ème expérience est, par l'usage de (1) : -1,9795, valeur bien approchée par le développement de q.

3.3 Une première relation différentielle de φ en tant que fonction de la fonction q.

Considérons l'intensité d'implication φ comme fonction de $q(a,\bar{b})$:

$$\varphi(q) = \frac{1}{\sqrt{2\pi}}\int_q^{\infty} e^{-t^2/2}dt$$

On peut alors examiner comment φ(q) varie lorsque q varie au voisinage d'une valeur donnée (a,b), sachant comment q varie lui-même en fonction des 4 paramètres qui le déterminent. Par dérivation de la borne d'intégration, on obtient :

$$\frac{d\varphi}{dq} = -\frac{1}{\sqrt{2\pi}} e^{\frac{-q^2}{2}} < 0 \quad (6)$$

Ce qui confirme bien que l'intensité croît lorsque q décroît, mais la vitesse de croissance est précisée par la formule, ce qui permet d'étudier avec plus de précision les variations de φ. Puisque la dérivée de φ par rapport à q est toujours négative, la fonction φ est décroissante.

3.4 Exemple numérique

Reprenant les valeurs des occurrences observées dans les 2 expériences évoquées plus haut, on trouve pour q = -2,309, la valeur de l'intensité d'implication φ(q) est égale à 0,992. Appliquant la formule (6), la dérivée de φ par rapport à q est : - 0 ,02775 et l'accroissement négatif de l'intensité est alors : - 0,02775 . Δq = - 0,02775. 0,3272. L'intensité approchée au premier ordre est donc : 0, 992 - Δq soit 0,983.

Or le calcul réel de cette intensité est, pour q= -1,9795 , φ(q) = 0, 976

4 Examen d'autres indices

Contrairement à l'indice de base q et l'intensité d'implication qui mesure la qualité à travers une probabilité (cf. définition 3), les autres indices les plus courants se veulent eux-mêmes directement des mesures de qualité. Nous examinerons leurs sensibilités respectives aux variations des paramètres retenus dans la définition de ces

indices. Nous conservons les notations adoptées au paragraphe 2 et choisissons des indices qui sont rappelés dans (Gras et al, 2004), (Lenca et al., 2005) et (Gras et Couturier, 2010) .

4-1 L'indice de Loevinger

C'est un « ancêtre » des indices d'implication (Loevinger, 1947). Cet indice, noté H(a,b) varie de 1 à - ∞. Il est défini par :

$$H(a,b) = 1 - \frac{n n_{a\wedge\bar{b}}}{n_a n_{\bar{b}}}.$$

Sa dérivée partielle par rapport à la variable nombre de contre-exemples est donc :

$$\frac{\partial H}{\partial n_{a\wedge\bar{b}}} = -\frac{n}{n_a\, n_{\bar{b}}}$$

Ainsi l'indice d'implication est toujours décroissant avec $n_{a\wedge\bar{b}}$. S'il est "proche" de 1, l'implication est "presque" satisfaite. Mais cet indice présente l'inconvénient, ne se référant pas à une échelle de probabilité, de ne pas fournir de seuil de vraisemblance et d'être invariant dans toute dilatation de E, A , B et $A \cap \overline{B}$.

4-2 L'indice Lift

Il s'exprime par : $l = \frac{n.n_{a\wedge b}}{n_a n_b}$. Cette expression, linéaire par rapport aux exemples, peut encore s'écrire pour mettre en évidence le nombre de contre-exemples :

$$l = \frac{n.(n_a - n_{a\wedge\bar{b}})}{n_a.n_b}.$$

Pour étudier la sensibilité de l aux variations des paramètres, nous formons :

$$\frac{\partial l}{\partial n_{a\wedge\bar{b}}} = -\frac{1}{n_a.n_b}$$

Ainsi, la variation de l'indice Lift est indépendante de celle du nombre de contre-exemples. C'est une constante qui ne dépend que des variations des occurrences de a et de b. l décroît donc lorsque le nombre de contre-exemples croît, ce qui sémantiquement, est acceptable mais la vitesse de décroissance ne dépend pas de la vitesse de croissance de $n_{a\wedge\bar{b}}$.

4.3 L'indice MC

Il s'exprime ainsi : $m = \frac{n_a - n_{a\wedge\bar{b}}}{n_b.n_{a\wedge\bar{b}}} n_{\bar{b}}$. Remarquons qu'en étant indépendant de n, il n'a pas de sens statistique aussi intéressant.

Sa dérivé partielle par rapport au nombre de contre-exemples est :

$$\frac{\partial m}{\partial n_{a\wedge\bar{b}}} = -\frac{n_a.n_{\bar{b}}}{n_b}.\left(\frac{1}{n_{a\wedge\bar{b}}}\right)^2.$$

L'indice m décroît donc lorsque $n_{a\wedge\bar{b}}$ croît et la vitesse de décroissance est même plus rapide qu'avec le Lift et qu'avec l'indice q gaussien basique pour calculer l'intensité d'implication. Il ne résiste pas à l'instabilité du nombre de contre-exemples.

4.4 La confiance

Cet indice est le plus connu et le plus utilisé grâce à la caisse de résonance dont dispose une publication anglo-saxonne (Agrawal et al. 1993). Il est à l'origine de plusieurs autres indices communément employés qui n'en sont que des variantes satisfaisant telle ou telle exigence sémantique.. De plus, il est simple et s'interprète aisément et immédiatement.

$$c = \frac{n_{a\wedge b}}{n_a} \text{ ou } 1 - \frac{n_{a\wedge\bar{b}}}{n_a}$$

La première forme, linéaire par rapport aux exemples, indépendante de n_b, s'interprète comme une fréquence conditionnelle des exemples de b quand a est connu.

La sensibilité de cet indice aux variations des occurrences des contre-exemples se lit à travers la dérivée partielle :

$$\frac{\partial c}{\partial n_{a\wedge\bar{b}}} = -\frac{1}{n_a}$$

Par conséquent, la confiance croît quand $n_{a\wedge\bar{b}}$ décroît, ce qui est sémantiquement acceptable, mais la vitesse de variation est constante, indépendante de la vitesse de décroissance de ce nombre, des variations de n et de n_b. Le gradient de c ne s'exprime que par rapport à $n_{a\wedge\bar{b}}$ et à n_a :

$$\begin{pmatrix} -\frac{1}{n_a} \\ \frac{n_{a\wedge\bar{b}}}{{n_a}^2} \end{pmatrix}$$

Ceci peut apparaître comme une restriction du rôle des paramètres dans l'expression de la sensibilité de l'indice.

5 Champ de gradient, champ implicatif

5-1 Existence d'un champ de gradient

Nous revenons à l'indice implicatif $q(a,\bar{b})$. Considérons l'espace $\mathfrak{E}$ de dimension 4 où les points M ont pour coordonnées les paramètres relatifs aux variables binaires a et b, soit $(n, n_a, n_{b,} n_{a\wedge\bar{b}})$. $q(a,\bar{b})$ définit donc un champ scalaire en tant qu'application de R^4 dans R (plongement de N^4 dans R^4).

Pour que le vecteur grad q de composantes les dérivées partielles de q par rapport aux variables n, n_a, $n_{b,}$ $n_{a\wedge\bar{b}}$ définisse un champ de gradient - champ vectoriel particulier que nous appellerons aussi champ implicatif- il doit respecter le critère de Schwartz d'une différentielle totale exacte à savoir et par exemple :

$$\frac{\delta}{\delta n_{a\wedge\bar{b}}}\left(\frac{\delta q}{\delta n_b}\right) = \frac{\delta}{\delta n_b}\left(\frac{\delta q}{\delta n_{a\wedge\bar{b}}}\right)$$

et idem pour les autres variables prises deux à deux. Or on a bien par les formules (3) et (4) :

$$\frac{\delta}{\delta n_{a\wedge\bar{b}}}\left(\frac{\delta q}{\delta n_b}\right) = \frac{1}{2}\left(\frac{n_a}{n}\right)^{-1/2}\left(\frac{n_{\bar{b}}}{n}\right)^{-3/2} = \frac{\delta}{n_b}\left(\frac{\delta q}{\delta n_{a\wedge\bar{b}}}\right)$$

Ainsi, au champ de vecteurs C = (n, n_a, n_b, $n_{a\wedge\bar{b}}$) de $\mathfrak{E}$, dont nous préciserons la nature, correspond un champ de gradient G qui est dit dérivé du **potentiel** q. Le gradient grad q est donc le vecteur qui représente la variation spatiale de l'intensité du champ. Il est dirigé des faibles valeurs du champ aux valeurs plus élevées. En suivant le gradient en chaque point, on suit l'augmentation de l'intensité d'implication du champ dans l'espace et, en quelque sorte la vitesse avec laquelle elle change sous l'effet de la variation d'un ou plusieurs paramètres.

Par exemple, si l'on fixe 3 des paramètres n, n_a, n_b, $n_{a\wedge\bar{b}}$ donnés par la réalisation du couple (a,b), le gradient est un vecteur dont la direction indique la croissance ou la décroissance de q, donc la décroissance ou la croissance de |q| et par suite de φ en fonction des variations du 4ème paramètre. Nous l'avions indiqué plus haut en interprétant la formule (5).

5.2 Lignes de niveau ou équipotentielles

Une ligne ou une surface équipotentielle (ou de niveau) dans le champ C est une courbe de $\mathfrak{E}$ le long de laquelle ou sur laquelle un point variable M conserve la même valeur du potentiel q (par ex. lignes isothermiques sur le globe ou lignes de niveau d'une carte IGN). L'équation de cette surface[5] est, bien entendu :

$$q(a,\bar{b}) - \frac{n_{a\wedge\bar{b}} - \frac{n_a\, n_{\bar{b}}}{n}}{\sqrt{\frac{n_a\, n_{\bar{b}}}{n}}} = 0$$

[5] En géométrie différentielle, on dirait que cette surface est une variété (quasi)différentiable à bord, compacte, homéomorphe au pavé fermé des intervalles de variations des 4 paramètres. Notons que le point dont la composante n_b est égale à n (donc $n_{a\wedge\bar{b}} = 0$) est un point singulier (« catastrophique » au sens de René Thom) de la surface et q, le potentiel, n'est pas différentiable en ce point. Partout ailleurs, la surface est différentiable, les points sont tous réguliers. Si le temps, par exemple, paramètre les observations du processus dont (n, n_a, n_b, $n_{a\wedge\bar{b}}$) est une réalisation, à chaque instant correspond une fibre morphologique du processus représentée par une telle surface dans l'espace-temps..

Par suite, sur une telle courbe, le produit scalaire grad q. dM est nul. Ce qui s'interprète comme indiquant l'orthogonalité du gradient avec la tangente ou l'hyperplan tangent à la courbe, c'est-à-dire avec la ligne ou la surface équipotentielle. Dans une interprétation cinématique de notre problème, la vitesse de parcours de M sur la surface équipotentielle est orthogonale au gradient en M.

A titre d'illustration, relativement à un potentiel F dépendant de 2 variables seulement, la figure ci-dessous par exemple montre la direction orthogonale du gradient par rapport aux différentes surfaces équipotentielles le long desquelles le potentiel F ne varie pas mais passe de F=7 à F= 10.

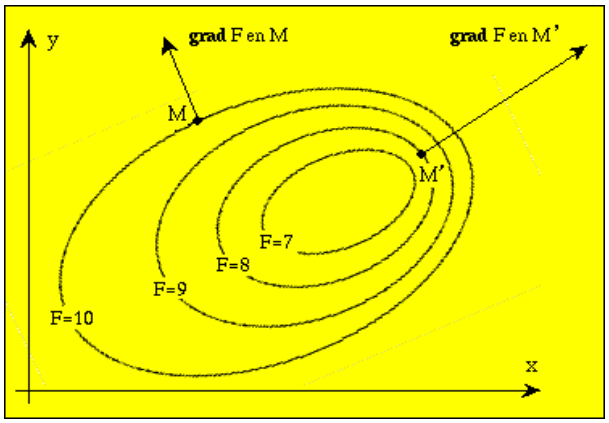


Figure 2

Il est possible dans le cas du potentiel q, de construire des surfaces équipotentielles comme ci-dessus (à deux dimensions pour la facilité de représentation). On peut comprendre que plus le champ est intense plus les surfaces sont serrées. Pour une valeur de q donnée, dans ce cas, on fixe 3 variables, par exemple n, n_a, n_b et une valeur de q compatibles avec les contraintes du champ. Soit : $n = 10^4$; $n_a = 1600 \leq n_b = 3600$ et $q = -2$ soit $|q| = 2$. On trouve alors $n_{a\wedge\bar{b}} = 528$ en utilisant la formule (1). Mais les points (10^4, 1600, 5100, 728) et (100, 25, 64, 3) appartiennent également à cette surface et la même courbe équipotentielle. Le point (10^4, 1600, 3600, 928) appartient à la courbe équipotentielle q=-3). En fait, sur toute cette surface, on obtient une sorte d'homéostasie de l'intensité d'implication.

L'expression de la fonction q de la variable $n_{a\wedge\bar{b}}$ permet de montrer qu'elle est convexe. Cette propriété prouve que le segment des points $t.M_1 + (1-t).M_2$, pour $t \in [0,1]$ qui joint deux points M_1 et M_2 de la même ligne équipotentielle est entièrement contenu dans sa convexité.

La figure ci-dessous représente, dans le champ implicatif deux surfaces équipotentielles voisines Σ_1 et Σ_2 correspondant à deux valeurs du potentiel q_1 et q_2. Au point M_1 le champ scalaire prend donc la valeur q_1. M_2 est l'intersection de la normale issue de M_1 avec Σ_2. Etant donné la direction du vecteur normal $\vec{n}$, la différence $\Delta = q_2 -$

q_1, variation du champ quand on passe de Σ_1à Σ_2 est alors égale à l'opposé de la norme du gradient de q en M_1 soit $\frac{\partial q}{\partial n}$, si n_a, n_b et $n_{a\wedge\bar{b}}$ sont fixés .

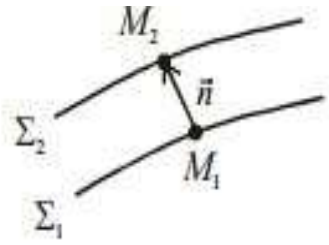


Figure 3

Ainsi, l'espace $\mathfrak{E}$ peut être comme feuilleté par des surfaces équipotentielles correspondant à des valeurs successives de q relativement aux cardinaux (n, n_a, n_b, $n_{a\wedge\bar{b}}$) que l'on ferait varier. Cette situation correspond à celle qui est envisagée dans la modélisation de l'ASI. Fixant n, n_a et n_b, on considère les ensembles aléatoires X et Y de mêmes cardinaux que A (n_a) et B (n_b) et dont le cardinal de $X \cap \bar{Y}$ suit une loi de Poisson ou une loi binomiale, suivant le choix du modèle. Les différents champs de gradient, véritables « lignes de force », qui leur sont associés sont orthogonaux aux surfaces définies par les valeurs correspondantes de Q. Ceci nous évoque, dans le cadre théorique du potentiel, la métaphore prémonitoire de « flux implicatif » que nous avons exprimée dans (Gras et al. 1996). Derrière cette notion nous pouvons imaginer un transport d'information d'intensité variable dans un univers causal. Nous illustrons cette métaphore avec l'étude des propriétés du cône implicatif à deux nappes (cf. Lahanier-Reuter et al, en publication ASI 8). De plus et intuitivement, l'implication a=>b est d'autant de bonne qualité que la surface équipotentielle C de la contingence recouvre des surfaces équipotentielles aléatoires dépendant de la variable aléatoire $Q(a,\bar{b})$.

Rappelons la relation qui unit le potentiel q à l'intensité d'implication $\varphi(a,b)$ définie par:

$$\varphi(a,b)= \frac{1}{\sqrt{2\pi}} \int_{q(a,\bar{b})}^{\infty} e^{-\frac{t^2}{2}} dt$$

Remarque 1

On constate que l'intensité est également invariante sur toute surface équipotentielle de ses propres variations. Les portions de surfaces engendrées par q et par φ sont même en correspondance biunivoque. En termes intuitifs, on peut affirmer que lorsque l'une « enfle » l'autre se « dégonfle.

Remarque 2

Notons une fois encore une particularité de l'intensité d'implication. Alors que les surfaces engendrées par les variations des 4 paramètres des données ne sont pas invariantes par une même dilatation des paramètres, celles associées aux indices citées

dans le § 4 sont invariantes et présentent une même forme géométrique indifférenciée.

6 Quelques simulations

6.1 Surfaces équipotentielles

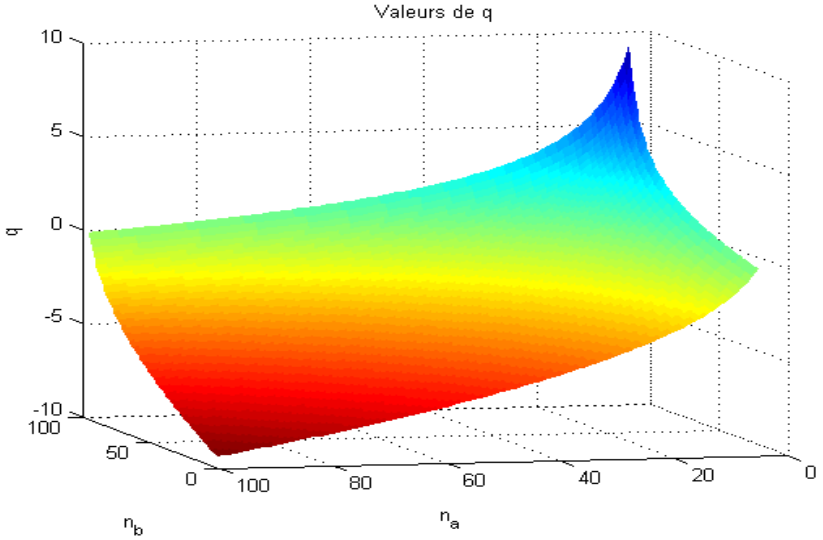
Afin de soutenir l'intuition et l'imagination, nous avons effectué quelques simulations de surfaces équipotentielles. Pour obtenir les figures suivantes avec Matlab, nous avons fixé n = 100, utilisé 3 valeurs de $n_{a\wedge\bar{b}}$: 5, 10 et 20 et avons fait varier n_a selon le 1er axe (dit des « x ») de $n_{a\wedge\bar{b}}$ à n et n_b selon le 2ème axe (dit des « y ») de 1 à $n - n_{a\wedge\bar{b}}$. Les courbes représentent donc les variations de q, indice d'implication selon l'axe vertical (dit des « z »). [6]

Sur ce type de figure en 3D, les valeurs affichées sur les axes, pour des commodités de représentation, peuvent être trompeuses. Sur la première par exemple les valeurs de n_a vont de 0 à 100 alors que n_a n'est bien entendu défini que de 5 à 100 puisque $n_{a\wedge\bar{b}} = 5$. De même, l'examen utile de la qualité de la règle a => b exige $n_a \leq n_b$.

Courbe 1 : $n_{a\wedge\bar{b}} = 5$

Les couleurs sont choisies pour signifier une certaine qualité de la règle :

- couleurs chaudes (rouge) pour les situations intéressantes où l'implication est satisfaite et d'autant plus que q est négatif,
- couleurs froides (bleu) où c'est le rejet de la règle a => b, voire la validité de la réciproque qui sont de plus en plus satisfaits et soulignés par l'intensité bleue des valeurs de q.



[6] Faisons une place à l'imagination : dans le cas présent où les variables et les paramètres associés ne prennent que des valeurs entières les points représentant les valeurs de q apparaissent comme des « petits pois » sur la « robe » équipotentielle.

Figure 4

Courbe 2 : $n_{a\wedge\bar{b}} = 20$

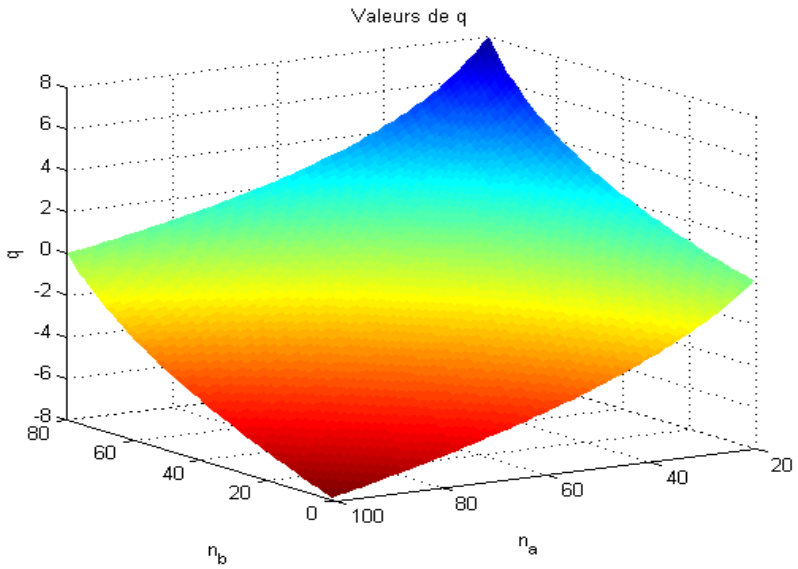


Figure 5

Remarquons l'étendue moins importante de la zone où l'implication est acceptable et, par contre, celle des rejets qui s'est amplifiée ainsi qu'un étalement plus marqué des valeurs indécises pour la règle.

6.2 Courbes équipotentielles de q pour différentes valeurs des contre-exemples

Courbes pour $n_{a\wedge\bar{b}} = 5$

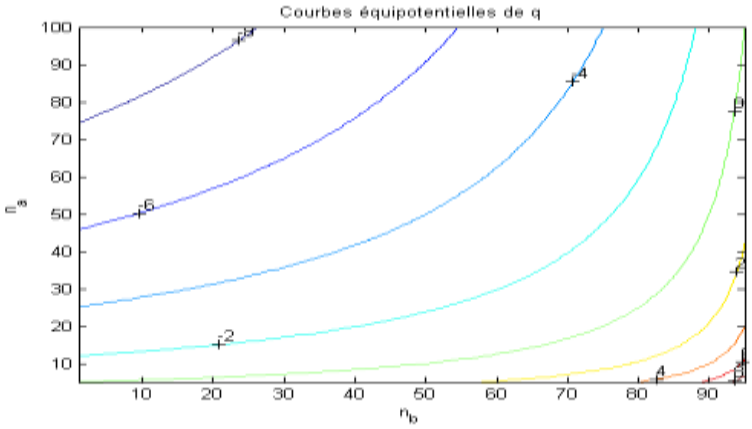


Figure 6

Ces courbes (n_b en abscisse et n_a en ordonnée) sont des projections planes des surfaces précédentes et limitées aux valeurs entières de q de -8 à 0 dans la figure 5 et de -6 à 0 pour la figure 6.

L'indice q est invariant tout le long de chaque arc de courbe. Par exemple, on trouve pour $n_a = 45$ et $n_b = 45$ (soit $n_{\bar{b}} = 55$), q = -3.9 (approximativement courbe q = -4). Mais aussi pour $n_a = 25$ et $n_b = 55$ (soit $n_{\bar{b}} = 45$), q = - 1.86.01$\cong -2$.

On peut remarquer que, pour des valeurs de n_a et de $n_{\bar{b}}$ assez élevées (donc n_b petit) les valeurs prises par q restent fortes plus « longtemps » que pour des valeurs opposées, ce que montre le resserrement des courbes pour des valeurs de q faibles conduisant à des incertitudes au sujet de la règle en raison de la décroissance de q.

Comparaison pour $n_{a\wedge\bar{b}}$ = 5 (courbes rouges) et $n_{a\wedge\bar{b}}$ = 10 (courbes vertes)

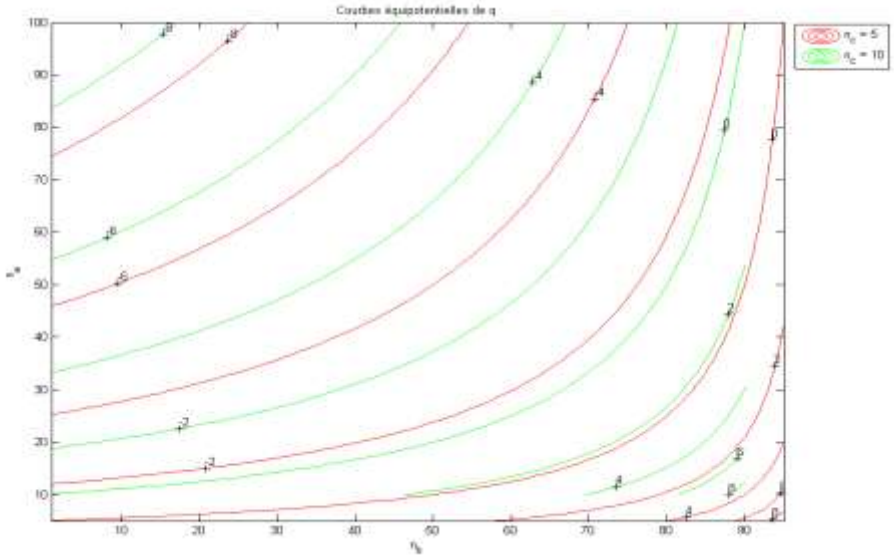


Figure 8

On notera le décalage des lignes équipotentielles rouges par rapport aux vertes où, pour une même valeur du couple (n_a, n_b), q est plus négatif sur les rouges que sur les vertes, donc accompagné d'une intensité d'implication de meilleure qualité.

- **Mise en évidence des vecteurs « gradient » du champ implicatif**.

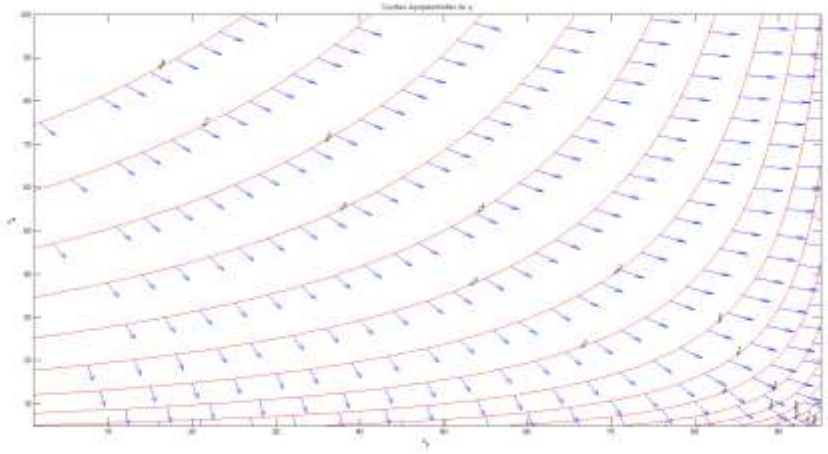


Figure 9

Si l'on fait maintenant apparaître les vecteurs « gradient » du champ implicatif normalisés et normaux aux lignes équipotentielles, on constate, métaphoriquement, une sorte de « respiration » du phénomène lié aux changements de valeurs des paramètres de la contingence. L'expansion de q, donc de l'intensité d'implication, est bien appréhendée par le sens du vecteur comme le laissaient prévoir les signes des dérivées partielles de q par rapport à n_a (négatif) et à n_b (positif) . Plus les paramètres croissent, pour un même nombre de contre-exemples, meilleures sont l'intensité et par conséquent notre confiance en la règle. Notons aussi que l'instabilité de q (courbes qui se « resserrent ») va croissante de gauche à droite comme on le remarquait dans le 6.2.

7 Conclusion

Sur la base des concepts qui ont conduit à la modélisation de la notion de quasi-implication en ASI, nous avons étudié l'indice primitif gaussien qui permet de quantifier la qualité de cette notion et, de ce fait, de qualifier notre agrément ou notre rejet des règles extraites d'un corpus de données binaires (et autres par extension). Notre étude s'est bien précisément centrée sur la sensibilité de l'indice d'implication aux variations des paramètres en jeu dans l'énonciation d'une règle. C'est ainsi que nous avons focalisé l'étude sur le champ scalaire défini par cet indice, puis sur un concept d'analyse, le gradient d'une telle fonction. Nous avons alors examiné les structurations de l'espace vectoriel donné par les paramètres, puis dégagé la notion de champ de gradient (dénommé ici champ implicatif) pour illustrer géométriquement, par des surfaces et des lignes équipotentielles, la sensibilité de l'indice d'implication. Ce pas-de-côté géométrique, par rapport aux approches statistiques et analytiques originelles, restitue à l'A.S.I. une vision dynamique du concept qui en enrichit la représentation mentale. A sa charge d'entretenir nos rêves car selon une formule de G. Bachelard *: « On ne peut étudier que ce que l'on a d'abord rêvé ».*

Références

Agrawal R., Imielinsky T. et Swami A,(1993), Mining association rules between sets of items in large databases, *Proc. of the ACM SIGMOD'93*, 207-216

Barbu E. (2003), *Hiérarchie cohésitive (ou implicative),* Mémoire de D.E .A. Extraction des Connaissances à partir des Données, Ecole Doctorale Informatique et Information pour la Société, Equipe COD, LINA, Université de Nantes.

Bernard J.-M. et Poitrenaud S, (1999) L'analyse implicative bayesienne d'un questionnaire binaire : quasi-implications et treillis de Galois simplifié", *Mathématiques, Informatique et Sciences Humaines*, n° 147, 25-46.

Cadot M., (2009), Graphe de règles d'implication statistique pour le raisonnement courant. Comparaison avec les réseaux bayesiens et les treillis de Galois, *Analyse Statistique Implicative, Une méthode d'analyse de données pour la recherche de causalités, sous la direction de Régis Gras, réd, invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse, p.223-250*

David J, Guillet F., Gras R. and Briand H. (2006): Conceptual hierarchies matching : an approach based on discovery of implication rules between concepts, *In Proc. ECAI*

2006, 17th European Conference on Artificial Intelligence, IOS Press, Riva del Garda, Italy.

D'Espagnat B. (1981), *A la recherche du réel*, Gauthier-Villars, Paris.

Fayyad U., Piatetsky-Shapiro G. and Smyth P. From Data Mining to Knowledge Discovery. In Advances In Knowledge Discovery and Data Mining, Fayyad U., Piatetsky-Shapiro G., Smyth P, and Uthurusamy R. eds, AAAI/MIT Press, 1-31.

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 1.

Gras R., Ag Almouloud S., Bailleul M., Larher A., Polo M., Ratsimba-Rajohn et Totohasina A (1996), *L'implication Statistique*, Collection Associée à Recherches en Didactique des Mathématiques, Grenoble : La Pensée Sauvage.

Gras R., Kuntz P. et Briand H., (2001b), Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, *Mathématiques et Sciences Humaines*, n° 154-155, 9-29.

Gras R., Couturier R.., Blanchard J., Briand H., Kuntz P., Peter P., (2004), Quelques critères pour une mesure de qualité de règles d'association. Un exemple : l'implication statistique, *Mesures de qualité pour la fouille de données, RNTI-E-1, Cépaduès –Editions*, 3-32.

Gras R., David J., Guillet F., Briand H. (2007). Stabilité en A.S.I. de l'intensité d'implication et comparaisons avec d'autres indices de qualité de règles d'association, *Proceedings atelier « Qualité des données et des connaissances », EGC 07*, Namur

Gras R., Couturier R., (2010) Spécificités de l'Analyse Statistique Implicative (A.S.I.) par rapport à d'autres mesures de qualité de règles d'association, *Quaderni di Ricerca in Didattica - GRIM (ISSN on-line 1592-4424)*, Eds : J.C. Régnier, R.Gras, F.Spagnolo, B. Di Paola, Université de Palerme, *p.19-57*

Lenca P., Vaillant B., Meyer P., et Lallich S. (2007), Association Rule Interestingness Measures : Experimental and Theorical Studies, *F.Guillet and H. J.Hamilton eds, Studies in Computational Intelligence 43, Springer*, p. 51-76.

Loevinger J. (1947), "A systematical approach to the construction and evaluation of tests of ability", *Psychological Monographs*, n° 61, 1947, p. 1-49

Ritschard G., Marcellin S., Zighed D.A. (2009), Arbre de décision pour données déséquilibrées : sur la complémentarité de l'intensité d'implication et de l'entropie décentrée, *Une méthode d'analyse de données pour la recherche de causalités, sous la direction de Régis Gras, réd, invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse p.207-21.*

Sève L. (2005), *Emergence, complexité et dialectique*, Odile Jacob, Paris.

Thom, R (1980). *Paraboles et catastrophes*, Flammarion.

Ouvrages de référence

L'implication statistique. Nouvelle méthode exploratoire de donnée, sous la direction de R.Gras, et la collaboration de S. Ag Almouloud, M. Bailleul, A. Larher, M. Polo, H. Ratsimba-Rajohn, A.Totohasina, La Pensée Sauvage, Grenoble (1996)

Mesures de Qualité pour la Fouille de Données, H.Briand, M.Sebag, R.Gras er F.Guillet eds, RNTI-E-1, Cépaduès, 2004

Quality Measures in Data Mining , F.Guillet et H.Hamilton eds, Springer, 2007,

Statistical Implicative Analysis, Theory and Applications, R.Gras, E. Suzuki, F. Guillet, F. Spagnolo, eds, Springer, 2008.

Analyse Statistique implicative.Une méthode d'analyse de données pour la recherche de causalités, sous la direction de Régis Gras, réd. invités R. Gras, J.C. Régnier, F. Guillet, Cépaduès Ed. Toulouse, 2009.

Teoria y Aplicaciones del Analisis Estadistico Implicativo, Eds : P.Orus, L.Zemora, P.Gregori, Universitat Jaume-1, Castellon (Espagne), ISBN : 978-84-692-3925-4, 2009..

L'Analyse Statistique Implicative : de l'exploratoire au confirmatoire. Eds : J.C. Régnier, Marc Bailleul, Régis Gras, Université de Caen, ISBN : 978-2-7466-5256-9, 2012

L'analyse statistique implicative, Méthode exploratoire et confirmatoire à la recherche de causalités, sous la direction de Gras R., eds Gras R., Régnier J.-C., Marinica C., Guillet F., Cépaduès Editions, 522 pages, ISBN 978.2.36493.056.8, 2013.