# Concordance entre deux Partitions : quelques propositions et expériences

Gilbert Saporta
Chaire de Statistique Appliquée et CEDRIC
CNAM
292 rue Saint Martin
75141 Paris Cedex 03
saporta@cnam.fr

Genane Youness
Institut des Sciences Appliquées et Economiques
CNAM-Université Libanaise et CEDRIC
BP113 6175 Hamra
Beyrouth, Liban
YOUNESS1@TERRA.NET.LB

RÉSUMÉ. *Nous proposons une méthodologie pour étudier la distribution du critère de Rand, dans le cas où les deux partitions devraient concorder.Utilisant un modèle de profils latents, on simule des données que l'on classe selon deux groupes de variables.La distribution de l'indice de Rand et d'un indice adapté du test de Mac Nemar se révèle être bimodale.*

MOTS-CLÉS : *classes latentes, K-means, indices d'associations, tests statistiques,partitions*

## 1. Introduction

Quand on dispose de deux partitions effectuées sur les mêmes individus, par exemple avec deux jeux de variables, ou bien avec deux algorithmes, la question se pose naturellement de savoir si ces deux partitions sont en accord ou bien si elles différent significativement, en un sens à préciser. Une maniére d'aborder ce problème consiste à calculer un indice de concordance entre partitions et à définir une valeur critique au dessus ou en deça de laquelle on concluera que les deux partitions sont ou non concordantes. A l'indice bien connu de Rand, nous proposons d'ajouter un indice inspiré du test de Mac Nemar .

Il faut alors connaitre, au moins approximativement, la distribution de probabilités de ces indices. Mais sous quelle hypothèse ? Cette question ne semble curieusement pas avoir été traitée dans la littérature, en tous cas pas sous des hypothèses réalistes [SAP 97]. En effet les rares travaux connus et récents [IDR 00], concernent la distribution de l'indice de Rand sous l'hypothèse d'indépendance. Or cette hypothèse n'est évidemment pas pertinente pour la question posée, car la non-indépendance ne signifie nullement une forte concordance.

La difficulté est de conceptualiser une hypothèse nulle d'identité de deux partitions. On se trouve dans une situation voisine de celle où on voudrait tester que deux variables numériques sont identiques : or si $\rho=1$, on sait bien que $r=1$ et on n'a donc pas de test utile de l'hypothèse nulle qui se trouve rejetée dès que $r<1$ .

Notre approche va consister à étudier la distribution de l'indice de Rand et de celui dérivé de Mac Nemar en engendrant par simulation des partitions qui devraient être proches car issues d'un même modèle sous-jacent : nous utiliserons pour cela un modèle de classes latentes régissant la distribution de p variables. On partage ensuite arbitrairement les p variables en deux groupes et comparons les partitions engendrées par une méthode classique de nuées dynamiques sur chacun de ces groupes de variables.

## 2. Indices d'accord entre partitions

### 2.1. *Notations*

$V_1$ et $V_2$ sont deux partitions des mêmes individus (ou deux variables qualitatives). On supposera que le nombre de classes k est le même. **N** désigne le tableau de contingence associé, $\mathbf{K_1}$, $\mathbf{K_2}$ les tableaux disjonctifs associés à $V_1$ et $V_2$; On a : $\mathbf{N} = \mathbf{K_1}\mathbf{K_2}$ .
Chaque partition $V_k$ peut aussi être caractérisée par son tableau de comparaisons par paires de taille nxn à valeur {0, 1} et dont le terme général $c_{ij}^{k}$ est défini par :

$C_{ij}^{k}$ = 1 si i et j sont dans la même classe de $V_k$

=0 sinon

On a $C_1=K_1K_1'$
Lorsque l'on croise deux partitions, on va s'intéresser aux paires d'individus qui restent ou ne restent pas dans les mêmes classes, d'où les 4 types :

Type 1 : paires dans une même classe de $V_1$ et dans une même classe de $V_2$
Type 2 : paires séparées dans $V_1$ et séparées dans $V_2$
Type 3 : paires séparées dans $V_1$ et dans une même classe de $V_2$
Type 4 : paires dans une même classe de $V_1$ et séparées dans $V_2$

Les effectifs de ces 4 types sont notés a, b, c , d et on a a+b+c+d= n(n-1)/2. On notera également A=a+b (nombre total d'accords) et D=c+d (nombre de désaccords)

### 2.2. L'indice de Rand

L'indice brut de Rand (semblable à l'indice de Kendall) est le pourcentage global de paires en accord :

$$R = \frac{A}{C_n^2}$$

On peut montrer que :

$$A = C_n^2 + \sum_{u=1}^{p}\sum_{v=1}^{q} n_{ij}^2 - \frac{1}{2}\left[\sum_{u=1}^{p} n_{i.}^2 + \sum_{v=1}^{q} n_{.j}^2\right]$$

Nous n'utiliserons pas la variante d'Hubert et Arabie [HUB 85], qui revient à un centrage réduction de R mais celle de Marcotorchino [MAR 91] où on considère toutes les paires, y compris celles identiques:

$$R = \frac{2\sum\sum n_{ij}^2 - \sum n_{i.}^2 - \sum n_{.k}^2 + n^2}{n^2}$$

car elle aboutit à une expression très simple en termes de comparaisons par paires :

$$R = \frac{\sum_i \sum_j (c_{ij}^1 c_{ij}^2 + \bar{c}_{ij}^1 \bar{c}_{ij}^2)}{n^2}$$

avec $\bar{c} = 1 - c$

C'est avec cette formulation relationnelle qu'Idrissi [IDR 00] a étudié la normalité asymptotique de R sous l'hypothèse d'indépendance. Par exemple si les k classes sont équiprobables on trouve que $c_{ij}^{1}c_{ij}^{2} + \bar{c}_{ij}^{1}\bar{c}_{ij}^{2}$ suit une loi de Bernoulli de paramètre $1-\frac{2}{k}+\frac{2}{k^2}$,

On en déduit :

$$E(R)=1-\frac{2}{k}+\frac{2}{k^2}$$

A. Idrissi affirme ensuite que le coefficient de Rand empirique entre deux variables qualitatives à k modalités équiprobables calculées sur n observations suit asymptotiquement une loi normale de variance :

$$V(R)=\frac{1}{n^2}(1-\frac{1}{n})(1-\frac{2}{k}+\frac{2}{k^2})(\frac{2}{k}-\frac{2}{k^2})$$

Cette expression de la variance suppose l'indépendance des $c_{ij}$ , ce qui est inexact en raison des contraintes de transitivité ($c_{ik}=c_{ij}\ c_{jk}$) est n'est vraie qu'approximativement pour k grand (il n'y a même pas normalité asymptotique pour des partitions en deux classes).

**2.3. L'indice dérivé de Mac Nemar**

Le test de Mac Nemar est un test non-paramétrique bien connu utilisé pour vérifier l'égalité de deux proportions dans des échantillons appariés (par exemple pourcentage d'individus favorables à une certaine opinion avant et après une campagne).

| | |
|---|---|
| a | b |
| c | d |

Si a désigne le nombre d'individus qui ont gardé la même opinion favorable, avant et après , d le nombre d'individus qui ont gardé la même opinion défavorable, b et c les effectifs de ceux qui ont changé d'avis, la statistique de test correspondant à l'hypothèse nulle selon laquelle les changements d'opinion dans un sens ou l'autre sont équiprobables est :

$$Mc=\frac{b-c}{\sqrt{b+c}}$$

Mc suit approximativement une loi normale N(0,1) sous $H_0$.

En adaptant ce test à l'ensemble des paires d'individus, on a une nouvelle façon de mesurer la concordance entre deux partitions, qui revient à se demander si les paires qui sont séparées le sont par hasard. On montre facilement que :

$$Mc=\frac{\sum_u n_{i.}^2-\sum_v n_{.j}^2}{2\sqrt{\frac{1}{2}(\sum_u n_{i.}^2+\sum_v n_{.j}^2)-\sum_u\sum_v n_{ij}^2}}$$

Ici encore,les relations de transitivité entre paires contredisent l'hypothèse d'indépendance entre observations, ce qui peut remettre en cause la normalité asymptotique.

### 3. Le modèle de classes latentes

Il faut maintenant définir ce que l'on entend par « deux partitions sont proches » : notre approche consiste à dire que les individus proviennent d'une même partition commune, dont les deux partitions observées en sont des réalisations bruitées.

Le modèle de classes latentes est bien adapté à cette problématique pour engendrer des partitions. Notons qu'il a été utilisé récemment pour la recherche de partitions consensus par Green et Kreiger [GRE 99]. Plus précisément, comme nous utiliserons des variables observées quantitatives, il s'agit d'un modèle de profils latents selon la terminologie de Bartholomew et Knott [BAR 99].

| | **Variables latentes** | |
|---|---|---|
| **Variables observées** | **qualitatives** | **quantitatives** |
| **qualitatives** | classes latentes | traits latents |
| **quantitatives** | profils latents | analyse factorielle |

**Tableau 1.** *Les méthodes de variables latentes*

L'hypothèse de base est l'indépendance des variables observées conditionnellement aux classes latentes :

$$f(\mathbf{x}) = \sum_k p_k \prod_j f_k(x_j / k)$$

Les $p_k$ sont les proportions des classes et **x** est le vecteur aléatoire des variables observées dont les composantes $x_j$ sont indépendantes dans chaque classe. On sait que ce modèle souffre de problèmes sérieux d'identifiabilité, mais ici il n'est utilisé que pour engendrer des données et non pour estimer des paramètres. Il suffit alors de générer des distributions indépendantes dans chaque classe, après avoir tiré le numéro de classe de chaque observation selon une multinomiale de probabilités $p_k$.

Pour obtenir des partitions «proches », qui ne différent l'une de l'autre que de façon aléatoire, nous proposons de construire des échantillons artificiels issus d'un modèle à k classes latentes et décrits par p variables numériques, que l'on supposera par commodité normales, mais d'autres distributions sont bien sur possibles. On partage ensuite arbitrairement les p variables en deux groupes et on effectue deux partitions en k classes des n individus selon ces deux groupes de variables à l'aide d'une méthode classique (les k-means ou nuées dynamiques.) Normalement, ces deux partitions doivent être peu différentes, on calcule alors l'indice de Rand ou l'indice dérivé de Mac Nemar.

On obtient un échantillon de valeurs de R et de Mc, sous l'hypothèse de « partitions proches » en itérant N fois, ce qui permet d'étudier leur distribution.

L'algorithme se déroule de la façon suivante :

1. Tirage des effectifs des classes latentes selon une loi multinomiale M(n; $p_1$ , ..,$p_k$)
2. Pour chaque classe, tirage de p variables normales indépendantes
3. Obtenir une partition $P_1$ sur $p_1$ variables et $P_2$ sur les autres p-$p_1$ variables
4. Calculer les indices de Rand et de Mac Nemar

**4 . Résultats expérimentaux**

Nous avons appliqué la procédure précédente en nous limitant à 4 classes latentes équiprobables, 1000 individus, et 4 variables. Les paramètres des distributions normales ont été choisis de telle sorte que pour chaque variable j, la valeur absolue de la différence entre les moyennes de la distribution normale de deux classes différentes soit plus de 1.5 fois son écart type :

$$| m_{kj} - m_{k'j} | > 1.5\, \sigma_j \qquad \forall\, j = 1, 2, 3, 4 \text{ et } \forall\, k \text{ et } k' = 1, 2, 3, 4$$

Le nombre d'itérations N vaut 1000.

Faute de place nous ne présentons qu'un des résultats de nos simulations (effectuées avec le logiciel S-plus).

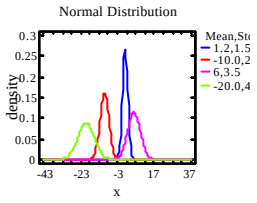
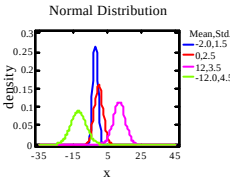
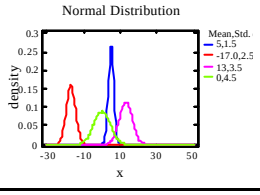
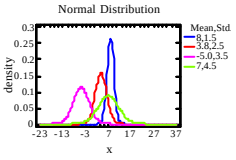
| | | |
|---|---|---|
| Classe 1 | X1 N( 1.2,1.5)<br>X2 N(-10,2.5)<br>X3 N(6,3.5)<br>X4 N(-20,4.5) |  |
| Classe 2 | X1 N( -2,1.5)<br>X2 N(0,2.5)<br>X3 N(12,3.5)<br>X4 N(-12,4.5) |  |
| Classe 3 | X1 N( 5,1.5)<br>X2 N(-17,2.5)<br>X3 N(13,3.5)<br>X4 N(0,4.5) |  |
| Classe 4 | X1 N(8,1.5)<br>X2 N(3.8,2.5)<br>X3 N(-5,3.5)<br>X4 N(7,4.5) |  |

**Tableau 2.** *Les distributions par classe*

Le graphique suivant dans le plan des deux premières composantes montre la répartition spatiale d'une des 1000 itérations :

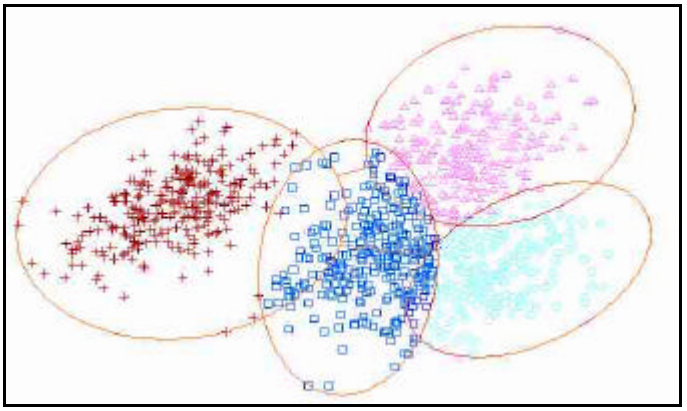


**Figure 1.** *Répartition des classes*

On effectue ensuite deux classifications par la méthode des k-means, l'une avec X1 et X2, l'autre avec X3 et X4, on calcule les indices de Rand et de Mac Nemar, et cela 1000 fois.

Les résultats obtenus sont assez surprenants : si on pouvait s'attendre pour R à une distribution non-normale, puisque la valeur théorique doit être proche de 1, rien ne nous laissait imaginer des distributions bi-modales :

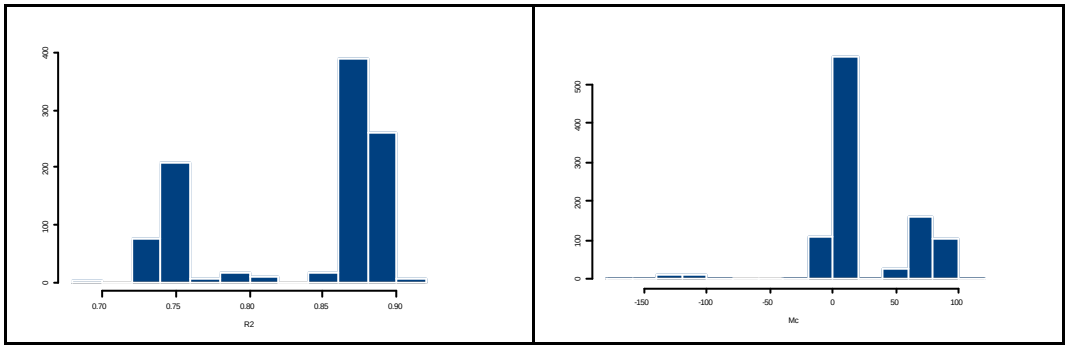



**Figure 2.** *Distribution des coefficients de Rand et Mac Nemar*

Ces configurations se sont reproduites à chaque jeu de simulations, pour des paramètres différents des lois normales. Notons que dans le cas précédent, toutes les valeurs observées du coefficient de Rand sont supérieures à 0.72, alors que l'espérance de R sous l'hypothèse d'indépendance est de 0.625, ce qui montre bien le caractère inadapté de celle-ci. Avec 1000 observations, on rejetterait l'indépendance si R>0.65 au risque 5% mais cela ne suffit pas pour montrer que les deux partitions sont « proches ».

On ne peut cependant proposer de seuil de signification pour chacun des coefficients, car les distributions dépendent non seulement du nombre de classes, mais de leur séparabilité qui est liée aux paramètres des distributions normales.

## 4. Discussion

L'utilisation d'un modèle de classes latentes a permis d'aborder le problème de la proximité de deux partitions. Celle-ci peut-être mesurée par l'indice de Rand ou celui dérivé de Mac Nemar. L'indice de Rand dans sa version utilisée, donne la même importance aux couples d'individus qui sont dans la même classe dans les deux partitions, qu'à ceux qui ne sont pas dans la même classe pour les deux partitions (accord «négatif »), ce qui est contestable.

La distribution de ces indices, qui est très différente de celle obtenue sous l'hypothèse d'indépendance, n'a été étudiée que par simulation et dans un cas très particulier. Il est clair que la distribution de Rand dépend de k, n et de la plus ou moins grande séparation des classes, etc. ce qui empêche de donner des bornes universelles. De plus le caractère bi-modal de cette distribution pose problème et nécessite encore des recherches.

## Bibliographie


[BAR 99] BARTHOLOMEW D.J., KNOTT M. *Latent Variable Models and Factor Analysis* , Arnold, London,1999.

[GRE 99] GREEN P., KREIGER A., « A Generalized Rand-Index Method for Consensus Clustering of Separate Partitions of the Same Data Base », *Journal of Classification*,16, 63-89,1999.

[HUB 85] HUBERT L., ARABIE P., « Comparing partitions », *Journal of Classification*, 2, 193-198, 1985.

[IDR 00] IDRISSI A. *Contribution à l'unification de Critère d'Association pour Variables Qualitatives*,Thèse de doctorat de l'Université de Paris 6, 2000.

[MAR 91] MARCOTORCHINO J.F., & EL AYOUBI, N., « Paradigme Logique Des Ecritures Relationnelles De Quelques Critères Fondamentaux D'Association », *Revue de Statistique Appliquée*, XXXIX , 2, 25-46, 1991 .

[SAP 97] SAPORTA, G. « Problèmes Posés par la Comparaison de Classifications Dans des Enquêtes Différentes », *53 ème session de l'Institut International de Statistique*, Istanbul, août 1997.