Sur des indices de comparaison de deux classifications

Genane Youness

CEDRIC-CNAM
BP 114661
Beyrouth, Liban
genane99@hotmail.com

Gilbert Saporta

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

RÉSUMÉ. On étudie la ressemblance entre deux partitions sur les mêmes individus à l'aide du coefficient de corrélation vectorielle RV et du kappa de Cohen (dans le cas où les partitions ont le même nombre de classes). On montre que le RV s'identifie à l'indice J de Janson et Vegelius. On étudie la distribution d'échantillonnage de ces indices pour des paires de partitions proches (issues d'un modèle de classes latentes) afin de donner des valeurs critiques sous des hypothèses réalistes.

MOTS-CLÉS : kappa de Cohen, corrélation vectorielle, indice de Janson et Vegelius, classification, classes latentes.

1. Introduction

Dans des travaux antérieurs [SAP 01, 02, 03], nous avons étudié la distribution du coefficient de Rand et d'indices voisins dans le but de comparer deux classifications provenant d'un même ensemble de données, afin de répondre aux questions suivantes : lors de deux enquêtes portant sur les mêmes individus, comment mesurer l'accord entre les deux classifications? Est-ce que les configurations de ces deux classifications se ressemblent ?

On présente ici les écritures logiques et relationnelles d'un indice obtenu à partir du coefficient de corrélation vectorielle RV introduit par P. Robert et Y. Escoufier [ROB 76] qui se révèle identique au coefficient J de S. Janson et J. Vegelius [JAN 82] ainsi que leurs distributions d'échantillonnage sous une hypothèse nulle d'absence de liaison.

Le coefficient kappa de Cohen, fournit une autre façon de mesurer l'accord entre deux partitions ayant le même nombre de classes, provenant d'un même échantillon. Cet indice, contrairement au précédent dépend de la numérotation des classes : on identifie la permutation des classes d'une des deux partitions en maximisant la valeur du kappa.

2. Notations

Soient P_1 et P_2 deux partitions des mêmes individus (ou deux variables qualitatives) à p et q classes. On notera K_1 et K_2 les tableaux disjonctifs associés et N le tableau de contingence croisant P_1 et P_2 de terme général n_{ij}. On a $N=K'_1K_2$

Lorsque l'on croise deux partitions, on s'intéresse également aux paires d'individus qui restent ou non dans les mêmes classes. On a en tout $n(n-1)/2$ paires d'individus.

A chaque partition P_k est associé un tableau relationnel C^k, de dimension nxn, dont le terme général $c^k_{ii'}$ est défini par :

$$c_{ii'}^{k} = \begin{cases} 1 \text{ si les deux individus i et i' sont dans la meme classe de la partition } P_k \\ 0 \text{ sinon} \end{cases}$$

On a $C^1 = K_1 K_1'$ et $C^2 = K_2 K_2'$

3. RV ou J

Le coefficient de corrélation vectorielle RV introduit par P. Robert et Y. Escoufier [ROB 76] permet de mesurer la ressemblance entre deux tableaux de données numériques X_1 et X_2 sur les mêmes observations en comparant les produits scalaires inter-individus associés aux deux tableaux .

Ces matrices de produits scalaires $X_i X'_i$ notées W_i sont de dimension *nxn* . Le coefficient RV est défini par :

$$RV(X_1, X_2) = \frac{\text{trace}(W_1 W_2)}{\sqrt{\text{trace}(W_1^2)\,\text{trace}(W_2^2)}}$$

Les travaux de A. Lazraq et R.Cleroux [LAZ 01,02] donnent la possibilité de tester des hypothèses concernant RV mais pour des données numériques.

Si on applique ce coefficient à deux tableaux disjonctifs K_1 et K_2 , on trouve :

$$RV(P_1,P_2) = \frac{\text{trace}(C^1 C^2)}{\sqrt{\text{trace}(C^1)^2\,\text{trace}(C^2)^2}} = \frac{\sum_{i,i'} (c_{ii'}^1)(c_{ii'}^2)}{\sqrt{\sum_{i,i'} (c_{ii'}^1)^2 \sum_{i,i'} (c_{ii'}^1)^2}}$$

Si RV est suffisamment grand, les classifications obtenues seront voisines.

En centrant les $c_{ii'}^{k}$ on retrouve la forme relationnelle de l'indice J de Janson et Vegelius établi par [IDR 00]:

$$J(P_1,P_2) = \frac{pq\sum\sum n_{ij}^2 - p\sum n_{i.}^2 - q\sum n_{.j}^2 + n^2}{\sqrt{[p(p-2)\sum n_{i.}^2 + n^2][q(q-2)\sum n_{.j}^2 + n^2]}} = \frac{\sum_{i,i'}(c_{ii'}^1 - \frac{1}{p})(c_{ii'}^2 - \frac{1}{q})}{\sqrt{\sum_{i,i'}(c_{ii'}^1 - \frac{1}{p})^2 \sum_{i,i'}(c_{ii'}^2 - \frac{1}{q})^2}}$$

4. Le kappa de Cohen

Introduit par [COH 60], le coefficient kappa est une mesure d'accord entre deux variables qualitatives pour des données appariées : pour deux partitions à même nombre de classes, il mesure l'écart à la diagonale du tableau de contingence :

$$K = \frac{n\sum_{i=1}^{k} n_{ii} - \sum_{i=1}^{k} n_{i.} n_{.i}}{n^2 - \sum_{i=1}^{k} n_{i.} n_{.i}}$$

La concordance observée P_0 est la proportion d'individus classés dans les cases diagonales de concordance du tableau de contingence, soit la somme des effectifs diagonaux divisés par la taille de l'échantillon *n* : $P_0 = \frac{1}{n}\sum_{i=1}^{k} n_{ii}$

La concordance aléatoire P_e est égale à la somme des produits des effectifs marginaux divisés par le carré de la taille de l'échantillon $P_e = \frac{1}{n^2}\sum_{i=1}^{k} n_{i.} n_{.i}$. Le kappa exprime la différence relative entre la proportion d'accords observés P_o et la proportion d'accords aléatoires P_e qui est la valeur espérée, sous l'hypothèse nulle d'indépendance des variables, divisée par le complément à un de l'accord aléatoire.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

L'identification des classes est nécessaire pour utiliser le coefficient kappa, car quand on a à comparer deux partitions des mêmes individus obtenues par des méthodes de classification, la numérotation des classes est arbitraire : il est alors logique d'identifier les classes des partitions qui conduisent à une valeur maximale de κ . On prend alors la permutation des classes qui maximise le kappa d'où leur renumérotation.

5. Distributions d'échantillonnage

On utilise la méthodologie présentée en [SAP 02] pour étudier la distribution d'échantillonnage de ces deux indices pour des paires de partitions proches.

Rappelons ici que le but n'est pas d'étudier si les deux partitions sont indépendantes, mais si elles sont concordantes : la difficulté étant de formuler correctement l'hypothèse nulle de concordance.On procéde alors comme suit: à partir d'une partition initiale basée sur un modèle de classes latentes, on obtient deux partitions par une méthode classique type k-means obtenue en séparant les variables en deux blocs. On calcule alors les indices ci-dessus pour les deux partitions : en itérant le procédé on obtient par simulation la distribution d'échantillonnage de RV (ou J) et de kappa.

6. Application numérique

On a obtenu deux partitions de 1000 individus P_1 et P_2 à 4 classes chacune par la méthode des k-means selon deux groupes de variables

Le tableau de contingence croisant les deux partitions est :

1	2	3	4
248	0	0	2
1	198	27	9
2	6	43	202
0	58	192	12

On trouve une valeur de l'indice Kappa égale à 0.335, et une valeur de J (ou RV) égale à 0.648.

On réordonne ensuite les colonnes selon le kappa maximal (il y a 4! permutations) pour pouvoir identifier les classes de P_2 à celles de P_1 : la valeur maximale du kappa est de 0.787 obtenue en permutant les deux dernières colonnes d'où le tableau réordonné :

1	2	4	3
248	0	2	0
1	198	9	27
2	6	202	43
0	58	12	192

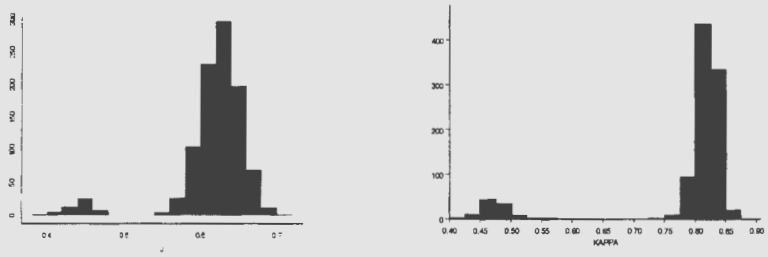

Figure 1 *: Distribution des coefficients de Janson et Vegelius et de kappa pour des partitions en 4 classes de 1000 individus avec 1000 itérations.*

Par simulation on trouve que le coefficient J varie entre 0.4 et 0.7. La valeur la plus fréquente est de 0.63 et la moyenne du coefficient J est égale à 0.617. Le coefficient kappa de Cohen varie entre 0.4 et 0.875, et est de moyenne 0.82. La bimodalité est due à la présence d'optimums locaux engendrés par la méthode des k-means. Dans l'exemple, on en déduit que les deux partitions sont suffisamment proches car les deux coefficients prennent des valeurs proches de la moyenne sous l'hypothèse de partitions identiques.

7. Conclusion

Nous avons montré l'identité des coefficients RV et J pour des partitions. J et le kappa de Cohen (mais ce dernier pour des nombres de classes identiques et après permutation) permettent de tester la similitude entre deux partitions en les comparant à leur distribution simulée dans le cas de données provenant d'une même partition « mère ».

Bibliographie

[COH 60] COHEN J., A coefficient of agreement for nominal scales., *Educ. Psychol. Meas.*, vol 20, 1960, p.27-46.

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

[LAZ 02] LAZRAQ, A., CLEROUX R., Inférence Robuste sur un indice de Redondance, *Revue de Statistique Appliquée*, vol. (4), p.39-54, 2002.

[JAN 82] JANSON S., VEGELIUS J., The J-index as a measure of association for nominal scale response agreement, *Applied psychological measurement*, vol. 16, 1982, p.243-250.

[ROB 76] ROBERT P., ESCOUFIER, Y. A unifying tool for linear multivariate statistical methods: the RV-coefficient, *Appl. Statist.*, vol. 25, 1976, p.257-65.

[SAP 01] SAPORTA G., YOUNESS G., Concordance entre deux partitions: quelques propositions et expériences, *in Actes des 8èmes rencontres de la SFC,* 2001, Pointe à Pitre.

[SAP 02] SAPORTA G., YOUNESS G., Comparing two partitions: some proposals and Experiments, *Proceedings in Computational Statistics edited by Wolfgang Härdle, Physica- Verlag*, 2002, Berlin.

[SAP 03] SAPORTA G., YOUNESS G., Une méthodologie pour la comparaison de partitions, *Revue de Statistique Appliquée*, à paraître.