Co-clustering sous différentes approches

Exercice 1 (10 points) : Dans une question à deux réponses, une mauvaise réponse est pénalisée : si la question vaut 1, une réponse fausse implique -0.5. Plusieurs bonnes réponses sont possibles pour certaines questions.

1. Le co-clustering est utilisé pour rechrcher	A –des blocs homogènes B –une partition de l'ensemble des lignes C –une partition de l'ensemble des colonnes	☐ ☐ ☐
2. Le co-clustering est particulièrement efficace sur des	A –tables binaires B –tables données continues comparables C –tables de contingences	☐ ☐ ☐
3. Le modèle de Bernoulli par blocs est un cas particulier du modèle multinomial	A – Vrai B – Faux	☐ ☐
4. L'algorithme Crobin permet de réaliser une classification sur des données de type	A – binaire B – continu C – qualitatif	☐ ☐ ☐
5. Quel le modèle de mélanges par blocs associé à Crobin ?	A –gaussien B –de Bernoulli C– mutinomial	☐ ☐ ☐
6. L'approche CML consiste à maximiser	A – la log-vraisemblance B – la log-vraisemblance classifiante C – ni l'une ni l'autre	☐ ☐ ☐
7. L'approche ML s'appuie sur l'algorithme	A – block EM B – block CEM C – double k-means	☐ ☐ ☐
8. La sparsité souvent présente dans les tables de co-occurrences est surmontée du fait que dans la classification croisée on travaille sur	A – des données positives B – des matrices intermédiaires non sparses	☐ ☐
9. Croeuc est un simple double k-means appliqué	A – sur des matrices intermédiaires réduites B – sur la matrice d'origine C – sur une matrice sparse	☐ ☐ ☐
10. Croeuc est basé sur une utilisation alternée de l'algorithme	A – spherical k-means B – k-means C – k-modes	☐ ☐ ☐
11. Dans le modèle de mélange de Bernoulli par blocs, si pour un block (kl) le paramètre α_{kl} est égal à 0.75 cela signifie que pour ce bloc on a	A – a_{kl} =1 et ε_{kl}=0.75 B – a_{kl} =0 et ε_{kl} =0.25 C – a_{kl} =1 et ε_{kl} =0.25	☐ ☐ ☐
12. Dans le modèle de Bernoulli par blocs, le paramètre ε_{kl} représente	A – le degré d'hétérogénéité B – le degré d'homogénéité C – la moyenne du bloc	☐ ☐ ☐
13. Le critère Croki2 est associé approximativement à un modèle	A – Bernoulli par blocs B – Poisson par blocs C – gaussiens par blocs	☐ ☐ ☐
14. En appliquant Croki2, il est judicieux de projeter les classes à l'aide des coordonnées d'une	A –ACP B –AFC C – AFCM	☐ ☐ ☐
15. Soit X la matrice de données de taille 1000x500 de type binaire, quel est le nombre de paramètres à estimer si l'on considère le modèle de Bernoulli à 3x2 composants dont les tailles sont supposées différentes.	……..	☐
16. Même question, mais cette fois ci- en considérant le modèle gaussien général	……..	☐
17. Même question mais en considérant le modèle de Poisson	……..	☐
18. Soit un modèle de mélange de Bernoulli par blocs. Si on suppose que ε_{11} =0.25, quelle la proportion des 1 dans le bloc (11)	A – 0.25 B – 0.75	☐ ☐
19. Dans le package {Blockcluster}, block EM est initialisé	A – par CAH B – au hasard C – par block CEM	☐ ☐ ☐
20. Appliqué sur une matrice de taille (n*p), Croeuc, Crobin et Croki2, cherchant un partitionnement en (g*m) blocs, consistent à travailler sur des matrices intermédiaires qui sont de taille	A – n*p B – n*m C – g*p	☐ ☐ ☐

Exercice 2 : Soit $X = (x_{ij})$ de taille *n* x *p*. Dans le modèle de Bernoulli par blocs [a_{kl}, ε] on a

$$f(x_{ij}, a_{kl}, \varepsilon) = \varepsilon^{|x_{ij}-a_{kl}|}(1-\varepsilon)^{1-|x_{ij}-a_{kl}|}$$

1. Définir la log-vraisemblance classifiante à maximiser ?
2. Ecrire les étapes de l'algorithme Block CEM.
3. Ecrire le code **R** permettant de lancer chacun des algorithmes à l'aide du package {blocluster}.

Exercice 3 : Sur le tableau 8x6 ci-dessous on applique l'algorithme Croeuc optmisant le critère

$$\sum_{i,j,k,l} z_{ik} w_{jl} \, (x_{ij} - \mu_{kl})^2 \qquad (1)$$

	1	2	3	4	5	6
A	0.9	2.9	2.9	3.0	3.0	2.9
B	0.9	2.9	2.9	3.1	3.0	2.9
C	2.8	0.9	1.0	1.1	1.0	0.9
D	3.0	1.0	0.9	1.0	0.9	0.9
E	2.8	0.9	0.9	1.1	0.9	1.0
F	0.9	3.0	2.9	3.0	3.1	2.9
G	2.9	0.8	0.9	1.0	1.1	1.0
H	1.1	3.0	2.9	3.0	3.1	3.1

1. A partir d'un couple de partition tiré au hasard $\mathbf{z}^{(0)}=(1,1,1,2,1,2,2,1)^T$ et $\mathbf{w}^{(0)}=(1,1,2,2,1,1)^T$, calculer la valeur du critère (1) après avoir réorganisé cette table.
2. Que représentent les termes suivants : $w_l = \sum_{j=1}^{6} w_{jl}$ et $z_k = \sum_{i=1}^{8} z_{ik}$,
3. Que représentent les termes suivants : $u_{il} = \frac{1}{w_l}\sum_{j=1}^{6} w_{jl}\, x_{ij}$ et $v_{kj} = \frac{1}{z_k}\sum_{i=1}^{8} z_{ik}\, x_{ij}$
4. Montrer que $\sum_{j=1}^{6} w_{jl}(x_{ij} - u_{il}) = 0$ et $\sum_{i=1}^{8} z_{ik}(x_{ij} - v_{kj}) = 0$
5. En déduire que minimiser $\sum_{i,j,k,l} z_{ik} w_{jl} \, (x_{ij} - \mu_{kl})^2$ revient à minimiser à
$$\sum_{i,k} z_{ik} \sum_{l} w_l (u_{il} - \mu_{kl})^2 = \sum_{j,l} w_{jl} \sum_{k} z_k (v_{kj} - \mu_{kl})^2$$
6. L'algorithme Croeuc alterne les deux minimisations. A la convergence on obtient $\mathbf{z}=(1,1,2,2,2,1,2,1)^T$, $\mathbf{w}=(1,2,2,2,2,2)^T$, calculer la valeur du critère (1).
7. Ecrire le code **R** permettant de réaliser ce co-clustering.

Exercice 4 : En appliquant BEM (Block EM) sur une matrice DATA on obtient les résultats et les figures ci-après.

1. De quel type de données s'agit-il : binaire, continu ou table de contingence ?
2. Quel est le modèle utilisé et le paramétrage employé ?
3. Quelle est la valeur du critère à la convergence ?
4. Quel est le bloc le plus homogène ? Quel est le bloc le moins homogène ?
5. Quel type de méthode de visualisation est approprié pour projeter les classes des lignes ?
6. Interpréter les distributions obtenues
7. Ecrire le code **R** permettant d'obtenir tous ces résultats.

```
*****************************************
Model Family : Gaussian Latent block model
Model Name : pik_rhol_sigma2kl
Co-Clustering Type : Unsupervised

Model Parameters..

Class Mean:
                [,1]       [,2]       [,3]
[1,] -0.0113171870  9.933649 -9.975353
[2,]  0.0008994504 -9.966978 10.031539

Class Variance:
          [,1]      [,2]     [,3]
[1,] 20.79533  9.854607 10.02493
[2,] 10.00552 19.958878 19.79938

Row proportions:  0.5 0.5
Column proportions:  0.28 0.32 0.4
Pseudo-likelihood:  -1.847052
```