Classification croisée de tableaux de contingence

Gérard Govaert

Heudiasyc, CNRS, Université de technologie de Compiègne

Travail mené avec M. Nadif, Université Paris-Descartes

6e Journées thématiques
Apprentissage Artificiel & Fouille de Données
29 Avril 2014

Plan

- **Les données**
- Mesures d'information
- Classification d'un tableau de contingence par approximation métrique
- Classification d'un tableau de contingence par modèle statistique
- Exemples

Lien entre deux variables qualitatives

- Lien entre la couleur des yeux et la couleur des cheveuxA
- Exemple issu du livre de D. Schwartz (Méthodes statistiques à l'usage des médecins et des biologistes)
- Echantillon de 124 personnes
- Deux variables qualitatives : couleur des yeux (3 modalités) et couleur des cheveux (4 modalités)

	yeux	cheveux
1	bleu	blond
2	vert	brun
3	bleu	noir
4	marron	roux
5	marron	brun
...	...	...
124	vert	blond

Lien entre deux variables qualitatives

- Lien entre la couleur des yeux et la couleur des cheveuxA
- Exemple issu du livre de D. Schwartz (Méthodes statistiques à l'usage des médecins et des biologistes)
- Echantillon de 124 personnes
- Deux variables qualitatives : couleur des yeux (3 modalités) et couleur des cheveux (4 modalités)

	yeux	cheveux
1	bleu	blond
2	vert	brun
3	bleu	noir
4	marron	roux
5	marron	brun
...	...	...
124	vert	blond

C. / Y.	blond	brun	noir	roux
bleu	25	9	3	7
vert	13	17	10	7
marron	7	13	8	5

Notations

I ensemble des lignes et J ensemble des colonnes

	1	...	j	...	d	
1	x_{1j}	...	x_{1j}	...	x_{1d}	$x_{1.}$
			$\vdots$	...		
i	x_{i1}	...	x_{ij}	...	x_{id}	$x_{i.}$
			$\vdots$	...		
n	x_{n1}	...	x_{nj}	...	x_{nd}	$x_{n.}$
	$x_{.1}$	...	$x_{.j}$	...	$x_{.d}$	N

Fréquences

	1	...	j	...	d	
1	p_{1j}	...	p_{1j}	...	p_{1d}	$p_{1.}$
			$\vdots$	...		
i	p_{i1}	...	p_{ij}	...	p_{id}	$p_{i.}$
			$\vdots$	...		
n	p_{n1}	...	p_{nj}	...	p_{nd}	$p_{n.}$
	$p_{.1}$	...	$p_{.j}$	...	$p_{.d}$	1

Distribution jointe

Exemple

	1	2	3	4	5	
1	5	4	6	1	0	16
2	6	5	4	0	1	16
3	1	0	1	7	5	14
4	1	1	0	6	5	13
5	4	5	3	4	5	21
6	5	4	4	3	4	20
	22	19	18	21	20	100

Fréquences

	1	2	3	4	5	
1	0.05	0.04	0.06	0.01	0.00	0.16
2	0.06	0.05	0.04	0.00	0.01	0.16
3	0.01	0.00	0.01	0.07	0.05	0.14
4	0.01	0.01	0.00	0.06	0.05	0.13
5	0.04	0.05	0.03	0.04	0.05	0.21
6	0.05	0.04	0.04	0.03	0.04	0.20
	0.22	0.19	0.18	0.21	0.20	1.00

Distribution jointe

Tableaux de comptage

- Issus directement du recueil de données
- Quelques exemples :
  - Tirage journalier moyen 2013 de quotidiens nationaux pour les grandes villes
  - Parcelles $\times$ espèces
  - documents $\times$ mots

Certains tableaux binaires

- Tableau de valeurs binaires
- Il ne s'agit du croisement de variables qualitatives binaires
- Tableau de présence absence
- Disymétrie entre les deux valeurs
- Exemple des données de référencement : clients x produits

Plan

- Les données
- **Mesures d'information**
- Classification d'un tableau de contingence par approximation métrique
- Classification d'un tableau de contingence par modèle statistique
- Exemples

Comment analyser ces données

- Prise en compte de la symétrie des deux ensembles mis en correspondance
- Exemple de l'analyse factorielle des correspondances
- Codage adapté + méthodes classiques (codage tf/idf + kmeans)
- Utilisation de mesures d'information ou d'association adaptées :
  - Nombreuses mesures
  - Choix de 2 mesures symétriques :
    - Coefficient du Φ^2 de Pearson
    - Information mutuelle

Coefficient du phi-2 de Pearson

- Nombreuses mesures d'association basées sur le χ^2 de contingence

$$\chi^2(\mathbf{x}) = \sum_{i,j} \frac{(x_{ij} - x_{i.}x_{.j}/N)^2}{x_{i.}x_{.j}/N}$$

- Coefficient du phi-2 de Pearson :

$$\Phi^2(P_{IJ}) = \frac{\chi^2(\mathbf{x})}{N} = \sum_{i,j} \frac{(p_{ij} - p_{i.}p_{.j})^2}{p_{i.}p_{.j}} = \sum_{i,j} \frac{p_{ij}^2}{p_{i.}p_{.j}} - 1$$

- Lien fort avec l'analyse factorielle des correspondances (AFC)

Information mutuelle

- Statistique du rapport de vraisemblance (likelihood ratio chi-squared)

$$G^2(\mathbf{x}) = 2 \sum_{i,j} x_{ij} \log \frac{x_{ij}}{x_{i.} x_{.j} / N}$$

- Par normalisation : information mutuelle

$$\mathrm{I}(P_{IJ}) = \frac{G^2(\mathbf{x})}{2N} = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i.} p_{.j}}$$

- Autre définition

$$\mathrm{I}(P_{IJ}) = H(P_I) + H(P_J) - H(P_{IJ})$$

où H est la fonction d'entropie

- Lien entre les 2 mesures

$$\mathrm{I}(P_{IJ}) = \frac{1}{2}\Phi^2(P_{IJ}) + o(\sum_{ij}(p_{ij} - p_{i.} p_{.j})^2)$$

Plan

- Les données
- Mesures d'information
- **Classification d'un tableau de contingence par approximation métrique**
- Classification d'un tableau de contingence par modèle statistique
- Exemples

Table de contingence associée à un couple de partitions

- $\mathbf{z}$ et $\mathbf{w}$ partitions en g classes et m classes des lignes et des colonnes de la table de contingence
- $\mathbf{z} = (z_1, \ldots, z_n)$ ou $\mathbf{z} = (z_{ik})$
- $\mathbf{w} = (w_1, \ldots, w_d)$ ou $\mathbf{w} = (w_{j\ell})$
- Nouvelle table de contingence $\mathbf{y^{zw}} = (y_{k\ell}^{\mathbf{zw}})$ à g lignes et m colonnes :

$$y_{k\ell}^{\mathbf{zw}} = \sum_{i,j} z_{ik} w_{j\ell} x_{ij}$$

- $K = \{1, \ldots, g\}$ et $L = \{1, \ldots, m\}$

Distributions associées à $\mathbf{z}$ et $\mathbf{w}$

- Distribution sur $K \times L$: $P_{KL}^{\mathbf{zw}} = (p_{k\ell}^{\mathbf{zw}})$ avec

$$p_{k\ell}^{\mathbf{zw}} = \frac{y_{k\ell}^{\mathbf{zw}}}{N} = \sum_{i,j} z_{ik}\, w_{j\ell}\, p_{ij} \qquad \forall (k,\ell) \in K \times L.$$

  - Les marges en lignes ne dépendent pas de $\mathbf{w}$: $p_{k.}^{\mathbf{z}}$
  - Les marges en colonnes ne dépendent pas de $\mathbf{z}$: $p_{.\ell}^{\mathbf{w}}$.

- Distribution sur $I \times J$: $Q_{IJ}^{\mathbf{zw}} = (q_{ij}^{\mathbf{zw}})$ avec

$$q_{ij}^{\mathbf{zw}} = p_{i.} p_{.j} \sum_{k,\ell} z_{ik}\, w_{j\ell} \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}}\, p_{.\ell}^{\mathbf{w}}}, \qquad \forall (i,j) \in I \times J$$

  - Les marges en lignes sont conservées : $q_{i.}^{\mathbf{z}} = p_{i.}$
  - Les marges en colonnes sont conservées $q_{.j}^{\mathbf{w}} = p_{.j}$

Exemples

- Exemple précédent

	1	2	3	4	5	
1	5	4	6	1	0	16
2	6	5	4	0	1	16
3	1	0	1	7	5	14
4	1	1	0	6	5	13
5	4	5	3	4	5	21
6	5	4	4	3	4	20
	22	19	18	21	20	100

Table de contingence

	1	2	3	4	5	
1	0.05	0.04	0.06	0.01	0.00	0.16
2	0.06	0.05	0.04	0.00	0.01	0.16
3	0.01	0.00	0.01	0.07	0.05	0.14
4	0.01	0.01	0.00	0.06	0.05	0.13
5	0.04	0.05	0.03	0.04	0.05	0.21
6	0.05	0.04	0.04	0.03	0.04	0.20
	0.22	0.19	0.18	0.21	0.20	1.00

Distribution jointe

- Partitions $\mathbf{z} = (1, 1, 2, 2, 3, 3)$ et $\mathbf{w} = (1, 1, 1, 2, 2)$

	1	2	
1	30.0	2.00	32.0
2	4.00	23.0	27.0
3	25.0	16.0	41.0
	59.0	41.0	100

Table de contingence agrégé $\mathbf{y}^{\mathbf{zw}}$

	1	2	
1	0.30	0.02	0.32
2	0.04	0.23	0.27
3	0.25	0.16	0.41
	0.59	0.41	1.00

Distribution associée $P_{KL}^{\mathbf{zw}}$

Exemples

- Distributions P_{IJ}

	1	2	3	4	5	
1	0.050	0.040	0.060	0.010	0.000	0.160
2	0.060	0.050	0.040	0.000	0.010	0.160
3	0.010	0.000	0.010	0.070	0.050	0.140
4	0.010	0.010	0.000	0.060	0.050	0.130
5	0.040	0.050	0.030	0.040	0.050	0.210
6	0.050	0.040	0.040	0.030	0.040	0.200
	0.220	0.190	0.180	0.210	0.200	1.000

- Q_{IJ}^{zw}

	1	2	3	4	5	
1	0.056	0.048	0.046	0.005	0.005	0.160
2	0.056	0.048	0.046	0.005	0.005	0.160
3	0.008	0.007	0.006	0.061	0.058	0.140
4	0.007	0.006	0.006	0.057	0.054	0.130
5	0.048	0.041	0.039	0.042	0.040	0.210
6	0.045	0.039	0.037	0.040	0.038	0.200
	0.220	0.190	0.180	0.210	0.200	1.000

Exemples

- Tableaux des $\frac{p_{ij}}{p_{i.}p_{.j}}$

	1	2	3	4	5
1	1.42	1.32	2.08	0.30	0.00
2	1.70	1.64	1.39	0.00	0.31
3	0.32	0.00	0.40	2.38	1.79
4	0.35	0.40	0.00	2.20	1.92
5	0.87	1.25	0.79	0.91	1.19
6	1.14	1.05	1.11	0.71	1.00

- Tableaux des $\frac{q_{ij}}{q_{i.}p_{.j}}$ et des $\frac{p_{kl}}{p_{k.}p_{.\ell}}$

	1	2	3	4	5
1	1.59	1.59	1.59	0.15	0.15
2	1.59	1.59	1.59	0.15	0.15
3	0.25	0.25	0.25	2.08	2.08
4	0.25	0.25	0.25	2.08	2.08
5	1.03	1.03	1.03	0.95	0.95
6	1.03	1.03	1.03	0.95	0.95

	1	2
1	1.59	0.15
2	0.25	2.08
3	1.03	0.95

Mesures d'association du ϕ^2 associées à **z** and **w**

$$\Phi^2(P_{KL}^{\mathbf{zw}}) = \sum_{k,\ell} \frac{(p_{k\ell}^{\mathbf{zw}} - p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}})^2}{p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}}} \qquad \Phi^2(Q_{IJ}^{\mathbf{zw}}) = \sum_{i,j} \frac{(q_{ij}^{\mathbf{zw}} - p_{i.} p_{.j})^2}{p_{i.} p_{.j}}$$

Propriétés :

- $\Phi^2(P_{KL}^{\mathbf{zw}}) = \Phi^2(Q_{IJ}^{\mathbf{zw}})$
- $\Phi^2(P_{IJ}) - \Phi^2(Q_{IJ}^{\mathbf{zw}}) = \Phi^2(P_{IJ}) - \Phi^2(P_{KL}^{\mathbf{zw}}) = D_{\Phi^2}(P_{IJ}, Q_{IJ}^{\mathbf{zw}})$
- $D_{\Phi^2}(P_{IJ}, Q_{IJ}^{\mathbf{zw}}) = \sum_{i,j} p_{ij} \left(\frac{p_{ij}}{p_{i.} p_{.j}} - \frac{q_{ij}^{\mathbf{zw}}}{p_{i.} p_{.j}} \right)^2$: distance du Φ^2 entre les 2 distributions P_{IJ} et $Q_{IJ}^{\mathbf{zw}}$
- $\Phi^2(Q_{IJ}^{\mathbf{zw}}) \leq \Phi^2(P_{IJ}) \quad \text{ou} \quad \Phi^2(P_{KL}^{\mathbf{zw}}) \leq \Phi^2(P_{IJ})$

Mesures d'information I associées à **z** and **w**

$$\mathrm{I}(P_{KL}^{\mathbf{zw}}) = \sum_{k,\ell} p_{k\ell}^{\mathbf{zw}} \log \frac{p_{k\ell}^{\mathbf{zw}}}{p_{k.}^{\mathbf{z}} p_{.\ell}^{\mathbf{w}}} \text{ and } \quad \mathrm{I}(Q_{IJ}^{\mathbf{zw}}) = \sum_{i,j} q_{ij}^{\mathbf{zw}} \log \frac{q_{ij}^{\mathbf{zw}}}{p_{i.} p_{.j}}$$

Propriétés :

- $\mathrm{I}(P_{KL}^{\mathbf{zw}}) = \mathrm{I}(Q_{IJ}^{\mathbf{zw}})$
- $\mathrm{I}(P_{IJ}) - \mathrm{I}(Q_{IJ}^{\mathbf{zw}}) = \mathrm{I}(P_{IJ}) - \mathrm{I}(P_{KL}^{\mathbf{zw}}) = \mathrm{KL}(P_{IJ} || Q_{IJ}^{\mathbf{zw}})$
- $\mathrm{KL}(P_{IJ} || Q_{IJ}^{\mathbf{zw}}) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}^{\mathbf{zw}}}$: distance de Kullback-Leibler entre les deux distributions P_{IJ} et $Q_{IJ}^{\mathbf{zw}}$
- $\mathrm{I}(Q_{IJ}^{\mathbf{zw}}) \leq \mathrm{I}(P_{IJ}) \quad \text{or} \quad \mathrm{I}(P_{KL}^{\mathbf{zw}}) \leq \mathrm{I}(P_{IJ})$

Classification croisée d'une table de contingence

- Deux approches équivalentes :
  - Recherche de la distribution réduite $P_{KL}^{\mathbf{zw}}$ conservant au mieux l'information initiale, c.-à-d. maximisant

$$\Phi^2(P_{KL}^{\mathbf{zw}}) \quad \text{ou} \quad \mathrm{I}(P_{KL}^{\mathbf{zw}})$$

  - Approximation de la distribution initiale P_{IJ} par une une distribution $Q_{IJ}^{\mathbf{zw}}$ minimisant la différence entre les mesures d'information de 2 distributions :

$$\Phi^2(P_{IJ}) - \Phi^2(Q_{IJ}^{\mathbf{zw}}) \quad \text{ou} \quad \mathrm{I}(P_{IJ}) - \mathrm{I}(Q_{IJ}^{\mathbf{zw}})$$

- Finalement, le problème est de trouver $\mathbf{z}$ et $\mathbf{w}$ minimisant le critère

$$W_{\Phi^2}(\mathbf{z},\mathbf{w}) = D_{\Phi^2}(P_{IJ}||P_{KL}^{\mathbf{zw}}) = \Phi^2(P_{IJ}) - \Phi^2(P_{KL}^{\mathbf{zw}}) = \Phi^2(P_{IJ}) - \Phi^2(Q_{IJ}^{\mathbf{zw}})$$

ou le critère

$$W_{\mathrm{I}}(\mathbf{z},\mathbf{w}) = \mathrm{KL}(P_{IJ}||P_{KL}^{\mathbf{zw}}) = \mathrm{I}(P_{IJ}) - \mathrm{I}(P_{KL}^{\mathbf{zw}}) = \mathrm{I}(P_{IJ}) - \mathrm{I}(Q_{IJ}^{\mathbf{zw}})$$

Modification du critère par augmentation des paramètres (1)

- Technique consistant à augmenter le nombre de paramètres d'un critère tel que la valeur optimale du critère initial ne soit pas modifiée mais plus facile à calculer
- Exemple des k-means :
  - Remplacement du critère $g(\mathbf{z}) = \sum_{i,k} d^2(\boldsymbol{x}_i, \mathbf{g}_k)$ par le critère

$$\widetilde{g}(\mathbf{z}, \mathbf{a}) = \sum_{i,k} d^2(\boldsymbol{x}_i, \mathbf{a}_k)$$

  - Paramètre additionnel $\mathbf{a} = (\mathbf{a}_1, \ldots, \mathbf{a}_g)$
  - Optimisation alternée de ce nouveau critère : k-means
  - Après chaque calcul de a, on retrouve le critère initial
- Ici, ajout du paramètre $\boldsymbol{\delta} = (\delta_{k\ell})$, matrice de dimension (g, m) vérifiant $\sum_{k,\ell} p^{\mathbf{z}}_{k.} p^{\mathbf{w}}_{.\ell} \delta_{k\ell} = 1$
- $\delta_{k\ell}$: « centre » du bloc $k\ell$

Modification du critère par augmentation des paramètres (2)

- Nouvelle distribution $R_{IJ}^{\mathbf{zw}\boldsymbol{\delta}}$: $r_{ij}^{\mathbf{zw}\boldsymbol{\delta}} = p_{i.}p_{.j}\sum_{k,\ell} z_{ik}w_{j\ell}\delta_{k\ell}$
- Propriétés : distribution sur $I \times J$ et mêmes marges que P_{IJ} et $Q_{IJ}^{\mathbf{zw}}$
- Nouveaux objectifs : minimisation de

$$\widetilde{W}_{\Phi^2}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = \Phi^2(P_{IJ}) - \Phi^2(R_{IJ}^{\mathbf{zw}\boldsymbol{\delta}})$$

ou de

$$\widetilde{W}_{\mathrm{I}}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = \mathrm{I}(P_{IJ}) - \mathrm{I}(R_{IJ}^{\mathbf{zw}\boldsymbol{\delta}})$$

qui peuvent aussi s'écrire

$$\widetilde{W}_{\Phi^2}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = \sum_{i,j,k;\ell} z_{ik}w_{j\ell}\, p_{i.}p_{.j}\left(\frac{p_{ij}}{p_{i.}p_{.j}} - \delta_{k\ell}\right)^2$$

et

$$\widetilde{W}_{\mathrm{I}}(\mathbf{z},\mathbf{w},\boldsymbol{\delta}) = \sum_{i,j} p_{ij}\log\frac{p_{ij}}{p_{i.}p_{.j}} - \sum_{k,\ell} p_{k\ell}^{\mathbf{zw}}\log\delta_{k\ell}$$

Algorithmes associés

- Croki2 : Optimisation alternée sur $\mathbf{z}$, $\mathbf{w}$ et δ du critère $\widetilde{W}_{\Phi^2}(\mathbf{z}, \mathbf{w}, \boldsymbol{\delta})$
- Croinfo : Optimisation alternée sur $\mathbf{z}$, $\mathbf{w}$ et δ du critère $\widetilde{W}_{\mathrm{I}}(\mathbf{z}, \mathbf{w}, \boldsymbol{\delta})$

Algorithme Croki2

input : $\mathbf{x}$, g et m

initialisation : $\mathbf{z}$ et $\mathbf{w}$ au hasard ; $\delta_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{j\ell} p_{ij}}{\sum_i z_{ik} p_{i.} \sum_j w_{j\ell} p_{.j}}$

repeat

- $z_i = \text{argmin}_k \sum_{j,\ell} w_{j\ell} p_{.j} \left(\frac{p_{ij}}{p_{i.} p_{.j}} - \delta_{k\ell} \right)^2$;
- $\delta_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{j\ell} p_{ij}}{\sum_i z_{ik} p_{i.} \sum_j w_{j\ell} p_{.j}}$;
- $w_j = \text{argmin}_\ell \sum_{i,k} z_{ik} p_{i.} (\frac{p_{ij}}{p_{i.} p_{.j}} - \delta_{k\ell})^2$;
- $\delta_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{j\ell} p_{ij}}{\sum_i z_{ik} p_{i.} \sum_j w_{j\ell} p_{.j}}$;

until convergence

return $\mathbf{z}$ and $\mathbf{w}$

Algorithme Croinfo

input : $\mathbf{x}$, g et m

initialisation : $\mathbf{z}$ et $\mathbf{w}$ au hasard ; $\delta_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{j\ell} p_{ij}}{\sum_i z_{ik} p_{i.} \sum_j w_{j\ell} p_{.j}}$;

repeat

- $z_i = \text{argmin}_k \sum_\ell (\sum_j w_{j\ell} p_{ij}) \log \delta_{k\ell}$;
- $\delta_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{j\ell} p_{ij}}{\sum_i z_{ik} p_{i.} \sum_j w_{j\ell} p_{.j}}$;
- $w_j = \text{argmin}_\ell \sum_k (\sum_i z_{ik} p_{ij}) \log \delta'_{k\ell}$;
- $\delta_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{j\ell} p_{ij}}{\sum_i z_{ik} p_{i.} \sum_j w_{j\ell} p_{.j}}$;

until convergence

return $\mathbf{z}$ and $\mathbf{w}$

Plan

- Les données
- Mesures d'information
- Classification d'un tableau de contingence par approximation métrique
- **Classification d'un tableau de contingence par modèle statistique**
- Exemples

Principe général

- Conditionnellement à la connaissance de la classe k de i et ℓ de j,

$$x_{ij} \sim \mathcal{P}(\mu_i \nu_j \gamma_{k\ell}) \qquad \forall i, j$$

  μ_i : effets ligne, ν_j : effets colonne, $\gamma_{k\ell}$: effets blocs
- Modèle non identifiable : contraintes sur les effets
- Estimation des effets lignes par les marges $x_{i.}$ et $x_{.j}$
- Modèles dépendant de $\mathbf{z}$, $\mathbf{w}$, $\boldsymbol{\gamma} = (\gamma_{11,}, \ldots, \gamma_{gm})$

Statut des ensembles I et J

- I et J fixés :
  - Table de contingence construite à partir de deux variables qualitatives
  - Taille de l'échantillon : N
  - *Clustering block model*
  - Paramètres : $\mathbf{z}$, $\mathbf{w}$, γ
- I et J échantillons :
  - Exemple clients $\times$ vidéos
  - Taille de l'échantillon : n, d
  - Variables latentes : $\mathbf{z}$, $\mathbf{w}$
  - *latent block model*
  - Paramètres : $\boldsymbol{\pi}$, $\boldsymbol{\rho}$, $\boldsymbol{\gamma}$
- I échantillon :
  - Exemple documents $\times$ mots
  - Taille de l'échantillon : n
  - Variable latente : $\mathbf{z}$
  - *Mixture model*
  - Paramètres : $\boldsymbol{\pi}$, $\mathbf{w}$, $\boldsymbol{\gamma}$

Clustering block model

- Modélisation classsique d'une table de contingence :

$$X \sim \mathcal{M}(N, p_{11}, \ldots, p_{ij}, \ldots, p_{nd})$$

  caractérisée par la probabilité p_{ij} de chaque cellule.
- Modèles log-linéaires
- Ici : $p_{ij} = \mu_i \nu_j \gamma_{k\ell}$
- Estimation par le maximum de vraisemblance
  - Maximisation de $L(\mathbf{z}, \mathbf{w}, \gamma) = \sum_{k,\ell} y_{k\ell}^{\mathbf{zw}} \log \gamma_{k\ell} + C$
  - Algorithme Croinfo
- Estimation par la méthode du khi-deux minimum
  - Minimisation de $\sum_{i,j} \frac{(x_{ij} - N p_{i.} p_{.j} \sum_{k\ell} z_{ik} w_{j\ell} \gamma_{k\ell})^2}{N p_{i.} p_{.j}}$
  - Algorithme Croki2

Latent block model

- Modèle

$$f(\mathbf{x};\boldsymbol{\pi},\boldsymbol{\rho},\boldsymbol{\gamma}) = \sum_{(\mathbf{z},\mathbf{w})\in\mathcal{Z}\times\mathcal{W}} \prod_{i,j} \pi_{z_i}\rho_{w_j}\mathcal{P}(x_{ij};\mu_i\nu_j\gamma_{z_i w_j})$$

- Interprétation
  1. Génération des n étiquettes $z_1,\ldots,z_n$ suivant la distribution multinomiale $\mathcal{M}(1,\pi_1,\ldots,\pi_g)$
  2. Génération des d étiquettes $w_1,\ldots,w_d$ suivant la distribution multinomiale $\mathcal{M}(1,\rho_1,\ldots,\rho_m)$
  3. Génération des x_{ij} suivant la loi de Poisson $\mathcal{P}(\mu_i\nu_j\gamma_{z_i w_j})$
- Estimation par le maximum de vraisemblance : algorithme EM
- Approximation variationnelle : LBVEM
- Approximation classificatoire : LBCEM

Algorithme LBVEM

input : $\mathbf{x}$, g et m

initialisation : $\widetilde{\mathbf{z}}$ et $\widetilde{\mathbf{w}}$ partition au hasard ; $\pi_k \leftarrow \frac{\widetilde{z}_{.k}}{n}$, $\rho_\ell \leftarrow \frac{\widetilde{w}_{.\ell}}{d}$ and $\gamma_{k\ell} \leftarrow \frac{\sum_{i,j} \widetilde{z}_{ik}\widetilde{w}_{k\ell}x_{ij}}{\sum_i \widetilde{z}_{ik}x_{i.}\sum_j \widetilde{w}_{j\ell}x_{.j}}$;

repeat

- $\tilde{z}_{ik} \propto \pi_k \exp(\sum_\ell (\sum_j \widetilde{w}_{j\ell}x_{.j}) \log \gamma_{k\ell})$
- $\pi_k \leftarrow \frac{\widetilde{z}_{.k}}{n}$, $\gamma_{k\ell} \leftarrow \frac{\sum_{i,j} \widetilde{z}_{ik}\widetilde{w}_{k\ell}x_{ij}}{\sum_i \widetilde{z}_{ik}x_{i.}\sum_j \widetilde{w}_{j\ell}x_{.j}}$;
- $\tilde{w}_{j\ell} \propto \rho_\ell \exp(\sum_k (\sum_i \widetilde{z}_{ik}x_{i.}) \log \gamma_{k\ell})$
- $\rho_\ell \leftarrow \frac{\widetilde{w}_{.\ell}}{d}$ and $\gamma_{k\ell} \leftarrow \frac{\sum_{i,j} \widetilde{z}_{ik}\widetilde{w}_{j\ell}x_{ij}}{\sum_i \widetilde{z}_{ik}x_{i.}\sum_j \widetilde{w}_{j\ell}x_{.j}}$;

until convergence

return Paramètres $\boldsymbol{\pi}$, $\boldsymbol{\rho}$ and $\boldsymbol{\gamma}$

Algorithme LBCEM

input : $\mathbf{x}$, g et m

initialization :

. $\widetilde{\mathbf{z}}$ et $\widetilde{\mathbf{w}}$ partition au hasard ; $\pi_k \leftarrow \frac{z_{.k}}{n}$, $\rho_\ell \leftarrow \frac{w_{.\ell}}{d}$ and $\gamma_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{k\ell} x_{ij}}{\sum_i z_{ik} x_{i.} \sum_j w_{j\ell} x_{.j}}$;

repeat

. $z_i \leftarrow \text{argmax}_k \, \pi_k \exp(\sum_\ell (\sum_j w_{j\ell} x_{.j}) \log \gamma_{k\ell})$;

. $\pi_k \leftarrow \frac{z_{.k}}{n}$, $\gamma_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{k\ell} x_{ij}}{\sum_i z_{ik} x_{i.} \sum_j w_{j\ell} x_{.j}}$;

. $w_j \leftarrow \text{argmax}_\ell \, \rho_\ell \exp(\sum_k (\sum_i z_{ik} x_{i.}) \log \gamma_{k\ell})$;

. $\rho_\ell \leftarrow \frac{w_{.\ell}}{d}$ and $\gamma_{k\ell} \leftarrow \frac{\sum_{i,j} z_{ik} w_{j\ell} x_{ij}}{\sum_i z_{ik} x_{i.} \sum_j w_{j\ell} x_{.j}}$;

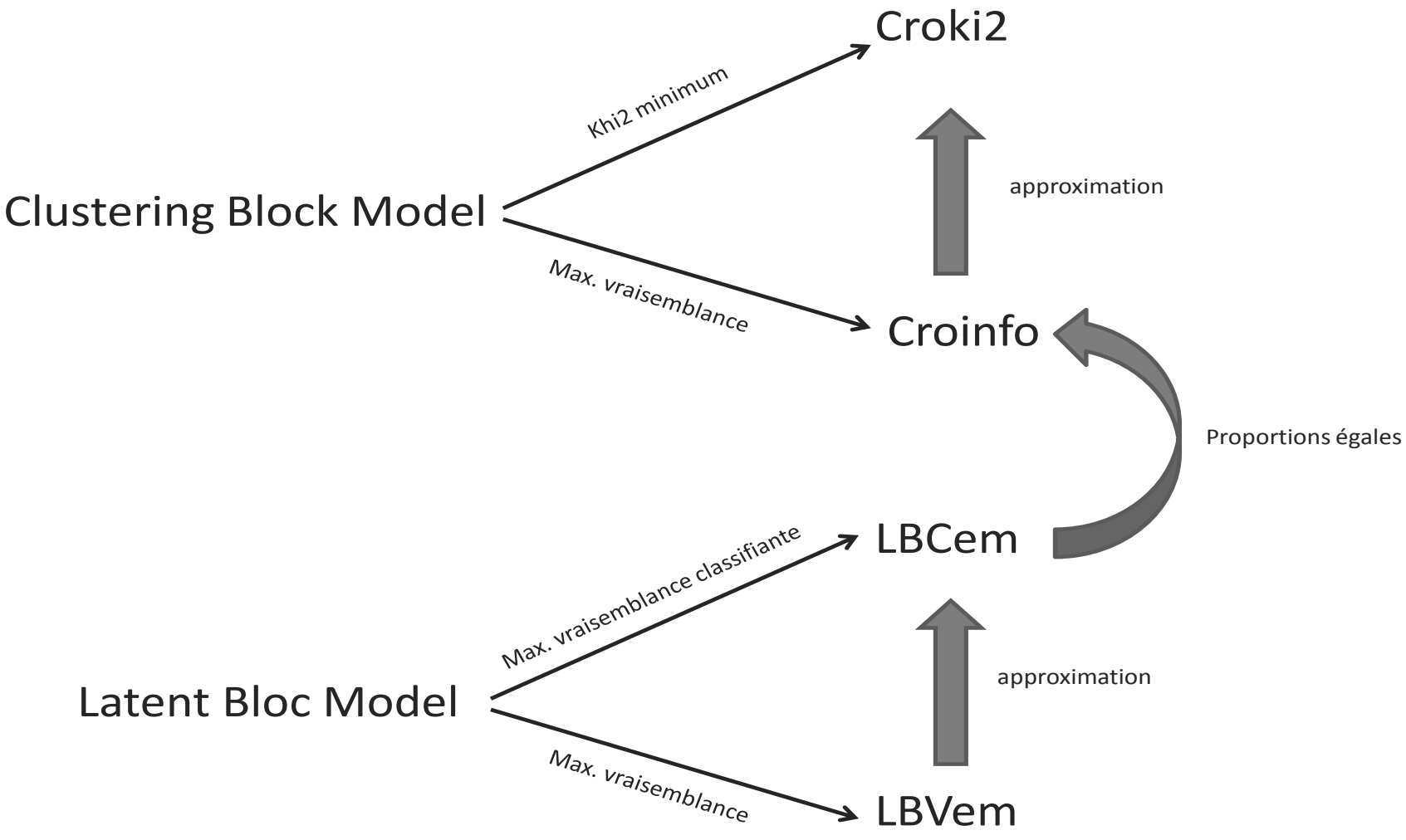
until convergence

return $\boldsymbol{\pi}$, $\boldsymbol{\rho}$ and $\boldsymbol{\gamma}$

Liens entre les critères

Algorithmes		Critères
Croki2	minimisation de	$\sum_{i,j,k;\ell} z_{ik} w_{j\ell}\, p_{i.} p_{.j} \left(\frac{p_{ij}}{p_{i.} p_{.j}} - \gamma_{k\ell} \right)^2$
Croinfo	minimisation de	$\sum_{i,j} p_{ij} \log \frac{p_{ij}}{p_{i.} p_{.j}} - \sum_{k,\ell} p_{k\ell}^{\mathbf{zw}} \log \gamma_{k\ell}$
LBCEM	maximisation de	$\sum_{i,j,k,\ell} z_{ik} w_{j\ell} (x_{ij} \log \gamma_{k\ell} - x_{i.} x_{.j} \gamma_{k\ell}) + \sum_{i,k} z_{ik} \log \pi_k + \sum_{j,\ell} w_{j\ell} \log \rho_\ell$
LBVEM	maximisation de	$\sum_{i,j,k,\ell} \tilde{z}_{ik} \tilde{w}_{j\ell} (x_{ij} \log \gamma_{k\ell} - x_{i.} x_{.j} \gamma_{k\ell}) + \sum_{i,k} \tilde{z}_{ik} \log \pi_k + \sum_{j,\ell} \tilde{w}_{j\ell} \log \rho_\ell - \sum_{i,k} \tilde{z}_{ik} \log \tilde{z}_{ik} - \sum_{j,\ell} \tilde{w}_{j\ell} \log \tilde{w}_{j\ell}$

Liens entre les algorithmes



Plan

- Les données
- Mesures d'information
- Classification d'un tableau de contingence par approximation métrique
- Classification d'un tableau de contingence par modèle statistique
- **Exemples**

Données Classic3 (1)

- Tableau de fréquences 3893 $\times$4303
- Lignes : 3893 résumés provenant de 3 bases de données :
  - *Medline* (1033) : médical
  - *Cisi* (1460) : informatique
  - *Cranfield* (1400) : systèmes aéronautiques
- Colonne : 4303 mots après filtrage
  - "petits mots"
  - mots peu fréquents
- Données « sparse » (99%)
- Application du modèle des blocs latents de Poisson
- Représentation visuelle des données par l'AFC cohérente avec les approches proposées

Données Classic3 (2)

- VEM et CEM donnent quasiment les mêmes résultats
- Comparaison avec Croinfo et un algorithme de classification croisée proposé par Dhillon (spectral clustering et théorie de l'information)
- 3 classes en ligne et 3 classes en colonne

	Med.	Cis.	Cra.
z_1	1007	3	2
z_2	25	1452	14
z_3	1	5	1384

LBVEM

	Med.	Cis.	Cra.
z_1	965	0	0
z_2	65	1458	0
z_3	3	2	1390

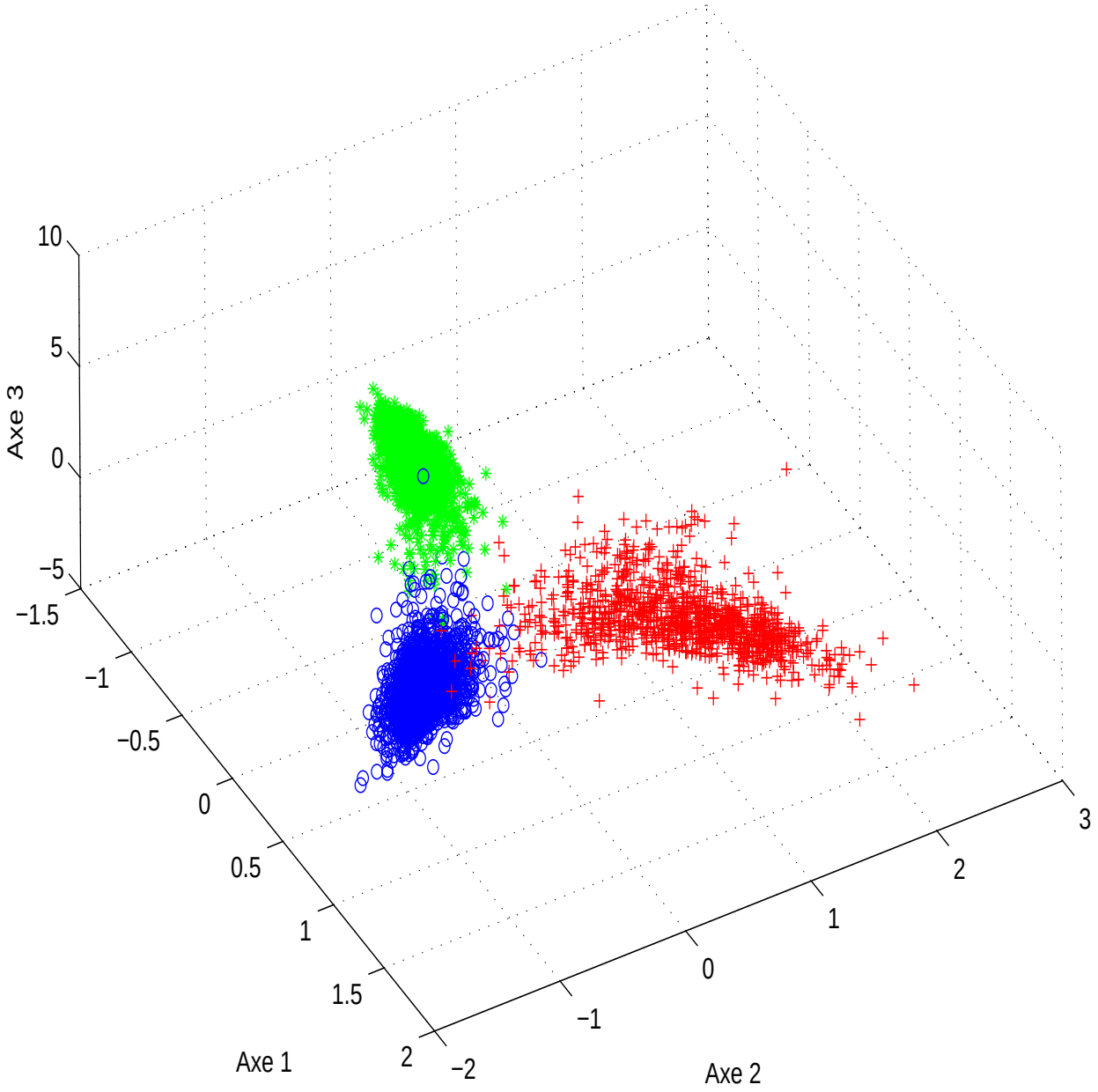
BSGP

	Med.	Cis.	Cra.
z_1	977	1	1
z_2	22	1454	15
z_3	34	5	1384

Croinfo

	LBVEM	BSGP	Croinfo
Nb de mal classés	50	70	78
% d'erreur	1.28	1.80	2.0

Données Classic3 (3)



Données Classic3 (4)

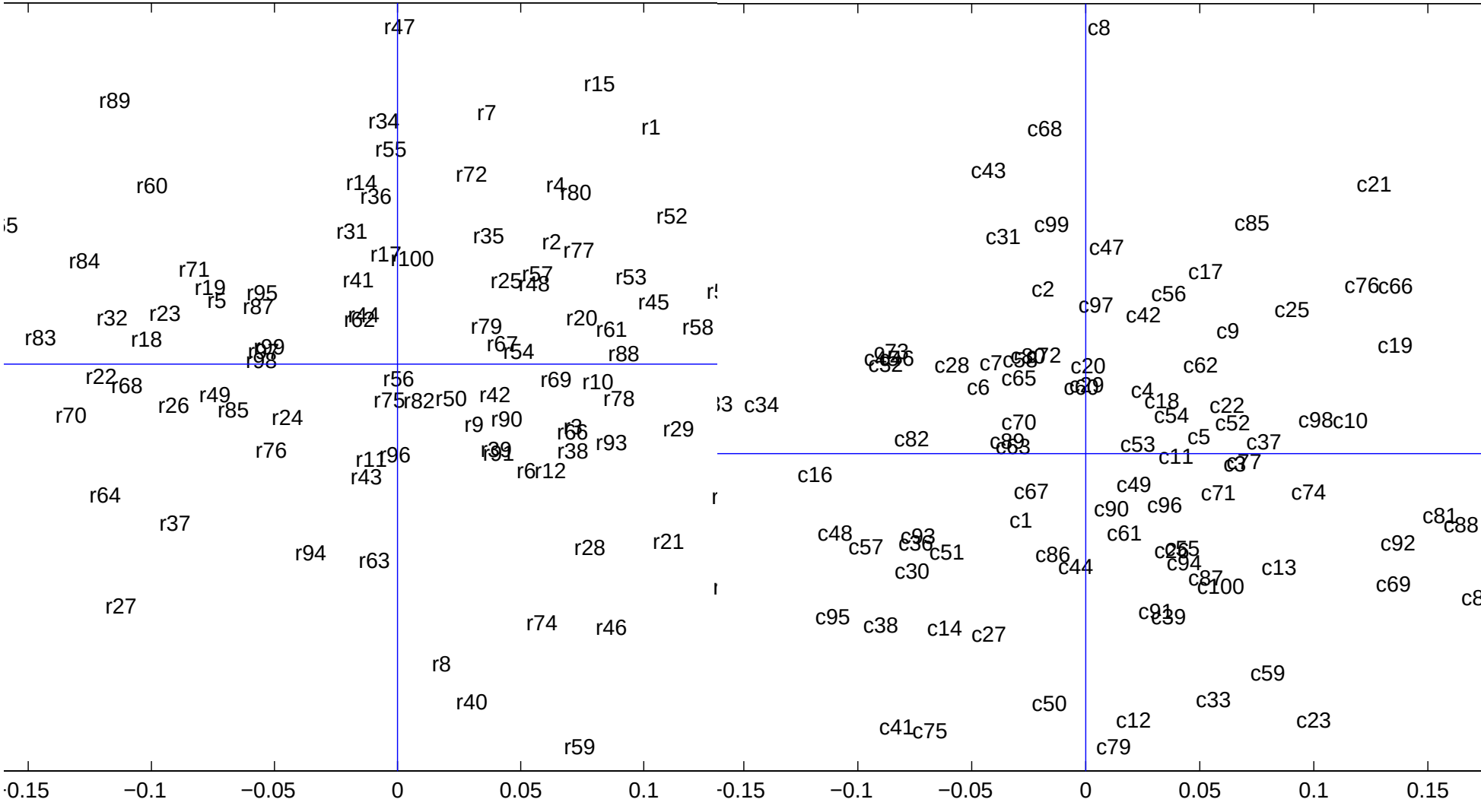
- Taux de sparsité 1%
- Taille du tableau : 134 Mo -> 2.9 Mo
- Tableau intermédiaire non sparse : 93 ko et 172 ko

Exemple montrant l'intérêt de la classification croisée

- Données simulées suivant le modèle des blocs latents de Poisson
- $n = d = 100$, $g = m = 3$
- Classes très mélangées
- AFC

	1	2	3	4	5	
Variance	0.007	0.006	0.006	0.006	0.005	0.005
Perc. of variance	4.428	3.537	3.505	3.355	3.134	2.943
Cumulated percentage	4.428	7.965	11.47	14.82	17.96	20.90

Représentation avec l'AFC



Taux d'erreur obtenus

	$(\mathbf{z}, \mathbf{w})$	$\mathbf{z}$	$\mathbf{w}$
Hasard	0.89	0.67	0.67
Théorique	0.48	0.28	0.28
VEM	0.50	0.33	0.26
EM		0.47	

Conclusion : avantages de la classification croisée

- Approche générique permettant de s'adapter à différentes situations
- Traitement symétrique des lignes et des colonnes ce qui semble souhaitable pour les tableaux binaires et les tableaux de contingence
- Modèles très parcimonieux
- Deux tableaux de travail de dimension $n \times m$ et $d \times g$:
  - Permet de traiter des données de grandes dimensions
  - Permet de traiter des données « sparse »
  - Dimensions « contrôlables »

Conclusion : problèmes ouverts

- Étude de l'identifiabilité des modèles
- Validation théorique de l'approximation variationnelle
- Initialisation des algorithmes et problème des classes vides
- Sélection de modèles :
  - Choix d'un modèle parcimonieux dans une famille de modèles
  - Choix des nombres de classes g et m
  - Extension des critères de vraisemblance pénalisée
  - Difficultés pour la taille des données : n, d, $n \times d$?

Bibliographie (1)

- Govaert G., Simultaneous Clustering of Rows and Columns, *Control and Cybernetics*, 24, 4, pp. 437-458, 1995.
- Govaert G. et Nadif M., Clustering with block mixture models, *Pattern Recognition*, 36, pp. 463-473, 2003.
- Govaert G. et Nadif M., An EM algorithm for the Block Mixture Model, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 4, pp. 643-647, 2005.
- Govaert G. et Nadif M., Fuzzy Clustering to estimate the parameters of block mixture models, *Soft Computing*, 10, 5, pp 415-422, 2006.
- Govaert G. et Nadif M., Clustering of contingency table and mixture model, *European Journal of Operational Research*, 183, pp. 1055-1066, 2007.

Bibliographie (2)

- Govaert G. et Nadif M., Block clustering with Bernoulli mixture models : Comparison of different approaches, *Computational Statistics and Data Analysis*, 52, 6, pp. 3233-3245, 2008.
- Govaert G. et Nadif M., Latent Block Model for Contingency Table, *Communications in Statistics - Theory and Methods*, 39, 3, pp. 416-425, 2010.
- Keribin C., Brault V., Celeux G. et Govaert G., Estimation and Selection for the Latent Block Model on Categorical Data, Statistics and Computing, 2014.