Université Lumière Lyon 2 - ICOM

Travail d'Etude et de Recherche

Co-clustering de données spatio-temporelles

Etudiants :
Mouhamadou Mansour LO
Mame Libasse MBOUP

Encadrant :
Julien JACQUES

13 mai 2022

Table des matières

1 Introduction

De nos jours, le monde qui nous entoure se remplit au fur et à mesure de l'explosion des nouvelles technologies de données brutes de plus en plus volumineux et complexe. Ces données peuvent provenir de diverses sources comme des capteurs d'IOT à hautes fréquences , des données publiques disponibles sur des sites comme Data-Gouv , des études de la population par l'INSEE . . . Ces données massives nécessitent d'être traitées et résumées pour qu'elles puissent être comprises et interprétées par des experts métiers. Dans le cas des données fonctionnelles multivariées, les méthodes classiques pour résumer cette information de manière non supervisée (clustering) semblent donner des résultats peu satisfaisants dus à une interprétation des moyennes de clusters compliqués.
Le co-clustering qui est une technique de classification consistant à réaliser une partition simultanée des lignes et des colonnes d'une matrice de données pourrait être un palliatif à notre problème.
Un récent papier sur le co-clustering de données fonctionnelles multivariées [1] de type "model-based" fournit une nouvelle approche et un package permettant de réaliser cette classification . Mais elle ne prend pas en compte l'existence potentielle de dépendances spatiales et temporelles dans les données.
Nous nous proposons dans ce qui suit d'étudier dans un premier temps le fonctionnement de cette nouvelle méthode puis dans un second temps de proposer une nouvelle implémentation de cet algorithme prenant compte de ces différentes dépendances.

Mots clés : co-clustering, multivariate functional data, funLBM, spatio-temporal data

2 Données Fonctionnelles Multivariées

Les données spatio-temporelles sont des données caractérisées par des attributs spatiaux (distance, direction, position, pression, température ...) et temporels (nombre d'occurrences, changements dans le temps, durée ...). Il s'agit de données qui subissent des évolutions dans l'espace et dans le Temps.

Ces données peuvent être collectées dans l'étude de divers domaines comme :

— en science du climat
— en science de la Terre
— en épidémiologie
— en neuroscience ...

Et peuvent servir à prédire entre autres des événements météorologiques, permet d'étudier la propagation des maladies, ou le suivi du trafic routier ...

Ces données sont par nature des données fonctionnelles multivariées qui elles, se définissent comme étant un ensemble de plusieurs fonctions de variables ou "Time series" décrivant un même individu.

Définition 01 :

* Soit un espace probabilisé $(\Omega, \mathcal{A}, \mathbb{P})$ avec (Ω un ensemble, $\mathcal{A}$ un tribu sur cet ensemble, $\mathbb{P}$ une mesure de probabilité sur $\mathcal{A}$)
* Soit $\mathcal{F}$, un espace de fonction (espace de Banach séparable)

Une variable fonctionnelle se définit comme étant : une variable aléatoire qui prend ses valeurs dans un espace vectoriel de dimension infini.

$$\mathcal{X} : \Omega \to \mathcal{F}$$

Définition 02 :

Une donnée fonctionnelle est une réalisation d'une variable fonctionnelle $\mathcal{X} : \Omega \times T \to \mathbb{R}$

- Dans le cas univarié , une donnée fonctionnelle est représentée par une unique courbe , $\mathcal{X}(t) \in \mathbb{R}, \forall t \in [0, T]$
- Dans le cas multivarié, on peut poser $\mathcal{X} = \mathcal{X}(t), \forall t \in [0, T]$, avec $\mathcal{X}(t) = (\mathcal{X}^1(t), ..., \mathcal{X}^P(t))^{'} \in \mathbb{R}^P, P \geq 2$

En réalité, l'expression fonctionnelle des courbes observées est inconnue, nous n'avons accès qu'aux observations discrètes mesurées en des temps précis .

Il faudrait donc reconstruire la forme fonctionnelle en exprimant les courbes dans un espace de dimension fini engendré par une base de fonctions .

Soit $\mathcal{X}_1, ..., \mathcal{X}_n$ un ensemble de n courbes $p - variees$ avec $\mathcal{X}_i = (\mathcal{X}_i^1, ..., \mathcal{X}_i^P)$. On suppose que chaque courbe est décrite par une combinaison linéaire de bases de :

$$\mathcal{X}_i^j(t) = \sum_{r=i}^{R_j} c_{ir}^j(\mathcal{X}_i^j)\phi_r^j(t)$$

Avec $i \in \{1, ..., n\}$, $j \in \{1, ..., p\}$, R_j : le nombre de bases de fonctions, $\phi_r^j(t)$: la base de fonction pour la j-ieme composante de la courbe multivariée, et c_{ir}^j : des coefficients.

En concaténant les coefficients c_{ir}^j en une matrice C et les bases de fonctions en $\Phi(t)$, $\mathcal{X}(t) = C\Phi^{'}(t)$

3 Qu'est-ce que le Co-clustering ?

Dans l'optique de comprendre et d'analyser un gros volume de données, on cherche à construire des groupes (classes ou clusters) similaires à partir d'un ensemble hétérogène d'objets. Pour s'approprier les propriétés de chaque cluster, une idée serait d'étudier un de ces représentants (la moyenne du groupe en guise d'exemple).

Sauf que chaque élément du même cluster a lui-même une quantité importante de variables qui le qualifie. Ce qui rend difficile l'analyse. D'où la nécessité de résumer parallèlement les variables ayant une distribution similaire en symétrie au regroupement des individus. Cette méthodologie est appelée le co-clustering (Goveart et al, 2013) [10]. Le co-clustering est alors une méthode récente réalisant simultanément un clustering des individus et des variables en classes homogènes.

Plusieurs techniques de co-clustering sont proposées à ce jour. Le biclustering, une autre appellation du co-clustering a été introduit par Hartigan [2] pour qui l'approche reposait sur une organisation au préalable de la structure du jeu de données.

Auparavant Good [3] proposé également une technique basée sur du clustering simultané des objets et des variables. De même Fisher [4] dans son étude, a défini un critère d'optimisation sur la façon simultanée de chercher à faire du clustering sur les dimensions d'une matrice de données en lignes et colonnes. Par contre cela n'a pas abouti à une solution résolvant la problématique.

Tryon et Bailey [5] quant à eux ont travaillé sur le co-clustering d'abord par le regroupement de l'ensemble des variables, par le biais de la matrice de corrélation, puis du regroupement des individus en se basant sur la mesure de distance entre les classes.

Nous pouvons citer également les travaux de Dubin et Champoux [6] dont la méthode combinait les variables en type et associait chaque individu aux types de variables afin de former une classification des individus, décrivant le choix d'une mesure de similarité et se contentant de mentionner la possibilité d'une classification des variables sans s'attarder sur la manière d'y parvenir.

Anderberg [7] a identifié le choix entre I (rows) et J(columns) parmi la liste des problèmes de classification. Il considère qu'il est raisonnable de classer les variables comme des individus. Il a même suggéré une approche itérative dans laquelle la classification est faite alternativement sur les individus et les variables jusqu'à les classifications des deux ensembles soient mutuellement "harmonieuses". Par ailleurs, il estime qu'une telle recherche offre simultanément "un potentiel considérable pour augmenter l'efficacité de la classification automatique.

Govaert [8] a développé un algorithme, en utilisant la statistique du carré du chi-deux ($\mathcal{X}^2$) comme mesure d'information, appelé CROKI2 pour rechercher simultanément des partitions de chaque ensemble, en minimisant la perte d'information due au regroupement des deux ensembles en classes. En étendant cette approche à des données binaires, continues et catégorielles, Govaert [9] a proposé les algorithmes CROBIN, CROEUC et CROMUL.

Deux familles de méthodes pour faire du co-clustering de lignes et de colonnes d'une matrice de données existent :

1. Les approches de type déterministe :
   (a) A scalable collaborative ltering framework based on co-clustering, George et al.
   (b) A generalized maximum entropy approach to bregman co-clustering and matrix approximation, Banerjee et al.
   (c) Penalized nonnegative matrix tri-factorization for coclustering, Wang et al.

2. Les approches de type model-based :
   (a) Model-Based Clustering and Classication for Data Science : With Applications in R, Bouveyron et al.
   (b) Latent Block Model , Co-Clustering (1st ed.) , Govaert et al.

Latent Block Model, a prouvé son efficacité pour le co-clustering sur différentes types de données :

— données continues (Nadif and Govaert, 2008)
— donéees nominales (Bhatia et al., 2014)
— données binaires (Laclau et al., 2017)
— ordinales (Jacques and Biernacki, 2018 ; Corneli et al., 2019)
— données functionnelles (Bouveyron et al., 2018 ; Chamroukhi and Biernacki, 2017).

Selosse et Jacques [11] ont présenté d'intéressants travaux sur le clustering de document. Leurs études ont porté sur le co-clustering de données avec à la fois des variables textuelles et continues par une extension du modèle des blocs latents.

On note par ailleurs Bouveyron, Jacques, Schmutz [1], qui ont proposé un nouveau modèle de co-clustering permettant de prendre en compte des courbes multivariées.

4 Jeu de données

Pour tester la pertinence du code que nous allons réalisé, il sera de mise de trouver des jeu de données fonctionelles multivariées avec éventuellement des variables dépendantes .

1. Datasets réels :
   (a) Pollution and météorologique data : Ce dataset proposé par AtmoSud est un jeu de données fonctionelles multivariées avec 1 variable temporelle , 1 variable spatiale et 5 variables de mesures météorologiques .
   (b) Ozone concentration data : Ce dataset proposé par State of Global Air est un jeu de données fonctionelles univarié avec 1 variable temporelle , 1 variable spatiale et 1 variable de mesure météorologiques.
2. Datasets simulés :
   un dataset simulé a été proposé dans l'article Co-clustering de courbes fonctionnelles multivariées, Schmutz et al. . ce dataset est fonctionnelle bivarié avec 100 lignes et 100 colonnes générées. Pour formule de génération Cf article.

D'autres dataset serons rajoutés au fur et à mesure des expériences .

5 Planifications

En terme de planification des expériences et du travail à réaliser, nous allons essayer dans les premières semaines de comprendre en détail le fonctionnement de l'algorithme MultiFunLBM proposé par l'article Co-Clustering of Multivariate Functional Data for the Analysis of Air Pollution in the South of France, Bouveyron et al. [1] et prendre en main l'outil FunLBM développé sous R associé à cet algorithme en réalisant plusieurs expériences sur des données simulées ou réelles .

En deuxième lieu , nous chercherons à étudier les méthodes et techniques présentées dans Model-based clustering for spatio-temporal data on air quality monitoring.,Cheam et al. et , Clustering spatial functional data, Vandewalle, pour prendre en compte les dépendances spatiales et temporelles de données fonctionnelles.

Puis nous allons chercher à travers divers expériences de rajouter au code de l'algorithme disponible dans FunLBM la possibilité de prendre en compte cette particularité.

6 Étude théorique de la problématique

Dans cette partie le travail consistera à discuter d'abord théoriquement du modèle **Latent Block Model** développé dans [1], ensuite apporter des idées sur comment étendre

ce modèle au cas où les **individus** sont **spatialement** dépendants et les **variables** temporellement **dépendantes**.

6.1 Latent Block Model

Latent Block Model(**LBM**) est un modèle de co-clustering qui **classifie** de manière conjointe les lignes et les colonnes d'une **matrice** de données.

Ce modèle **probabiliste** et **générative** met en place une double partition d'une matrice de données $X(n \times p)$: sur les **lignes** et sur les **colonnes**.

L'objectif ici est de résumer l'information de la matrice **X** en de petits sous-groupes appelés : **Blocks**.

Soit $z = (z_{ik})_{1 \leq i \leq n, 1 \leq k \leq K}$: la partition des **n** lignes (spatial) en **K groupes** et $w = (w_{jl})_{1 \leq j \leq p, 1 \leq l \leq L}$: la partition des **p** colonnes (temporal) en **L groupes** , tel que $z_{ik} = 1$ si la ligne **i** appartient au cluster-ligne **k** , 0 sinon (de meme pour w_{jl})

Un **block** se définit comme étant un ensemble de courbes multivariées qui appartient a un cluster-ligne et un cluster-colonne tel que $z_{ik}w_{jl} = 1$

6.1.1 Block Model

$$f(x;\theta) = \prod_{i,j} f(x_{ij};\alpha_{ziwj})$$

6.1.2 Mixture Model

$$f(x;\theta) = \prod_{i}(\sum_{k} \pi_k \prod_{j,l} \{f(x_{ij};\alpha_{kl}\}^{wjl})$$

$$\pi_k = P(z_{ik} = 1)$$

6.1.3 Probability Density Function (pdf)

$$f(x;\theta) = \sum_{(z,w) \in Z*W} P(z,w) f(x|z,w;\theta)$$

6.1.4 Les Hypothèses d'indépendances

Plusieurs hypothèses (4) d'indépendances sont faites :

— Clusters lignes et Clusters colonnes independants

$$P(z, w) = P(z)P(w)$$

— Clusters lignes indépendants et identiquement distribués

$$P(z) = \prod_i P(z_i) = \prod_{ik} \alpha_k^{z_{ik}}$$

— Clusters colonnes indépendants et identiquement distribués

$$P(w) = \prod_j P(w_j) = \prod_{jl} \beta_l^{w_{jl}}$$

— Pour une colonne et une ligne donnée , X_{ij} indépendants et identiquement distribués

$$P(x|z, w; \theta) = \prod_{ijkl} P(x_{ij}, \theta_{kl})^{z_{ik} w_{jl}}$$

De ce fait la forme finale du pdf de **x** est :

$$f(x; \theta) = \sum_{(z,w) \in Z*W} \prod_{ik} \alpha_k^{z_{ik}} \prod_{jl} \beta_l^{w_{jl}} \prod_{ijkl} P(x_{ij}, \theta_{kl})^{z_{ik} w_{jl}}$$

6.1.5 Mixture Model à partir de LBM

Exprimons le pdf conditionellemnt à **w** :

$$f(x|w) = \sum_z P(z) f(x|z, w)$$

$$f(x|w) = \sum_z P(z) \prod_{i,j} f(x_{ij}; \alpha_{ziwj})$$

$$f(x|w) = \sum_z \prod_i P(z_i) \prod_{i,j} f(x_{ij}; \alpha_{ziwj})$$

$$f(x|w) = \sum_z \prod_i (P(z_i) \prod_j f(x_{ij}; \alpha_{ziwj}))$$

$$f(x|w) = \prod_i \sum_k (\pi_k \prod_j f(x_{ij}; \alpha_{kwj}))$$

$$f(x|w) = \prod_i \sum_k (\pi_k \prod_{j,l} f(x_{ij}; \alpha_{kl})^{wjl})$$

Exprimons le pdf conditionellemnt à **z** :

$$f(x|z) = \sum_{w} P(w) f(x|z,w)$$

$$f(x|z) = \sum_{w} P(w) \prod_{i,j} f(x_{ij}; \alpha_{ziwj})$$

$$f(x|z) = \sum_{w} \prod_{j} P(w_j) \prod_{i,j} f(x_{ij}; \alpha_{ziwj})$$

$$f(x|z) = \sum_{w} \prod_{j} (P(w_j) \prod_{i} f(x_{ij}; \alpha_{ziwj}))$$

$$f(x|z) = \prod_{j} \sum_{l} (\pi_l \prod_{i} f(x_{ij}; \alpha_{zil}))$$

$$f(x|z) = \prod_{j} \sum_{l} (\pi_l \prod_{i,k} f(x_{ij}; \alpha_{kl})^{zik})$$

Exprimons le pdf conditionnellement à **z** et **w** :

$$f(x) = \sum_{(z,w) \in Z*W} P(z,w) f(x|z,w)$$

$$f(x) = \sum_{(z,w) \in Z*W} \prod_{i} P(z_i) \prod_{j} P(w_j) \prod_{i,j} f(x_{ij}; \alpha_{ziwj})$$

$$f(x) = \prod_{i,j} (\sum_{k} \sum_{l} \pi_{kl} P(x_{ij}, \theta_{kl}))$$

$$\pi_{kl} = P(z_{ik} w_{jl}) = 1$$

6.2 Framework proposé pour le clustering de données fonctionnelles spatiales

Soit le modele de melange **standard**

$$f(x) = \sum_{g=1}^{G} \pi_g f_g(x)$$

L'idée proposée par [16] est de **parametriser** les individus **spatialement** :

$$f(x|s;\theta) = \sum_{g=1}^{G} \pi_g(s,\beta) f_g(x, \theta_g)$$

Dans un même **cluster** (**Block** dans notre cas) , la distribution des individus est indépendante de la localisation.

La **dépendance spatiale** est capturée par $\pi_g(s, \beta)$.

Dans l'article proposé par **Cheam et al** , la **régression logistique multinomiale** est utilisée comme modèle pour $\pi_g(s, \beta)$

6.3 Régression Logistique Multinomiale

La Régression Logistique **Multinomiale** ou **Polytomique** est une méthode de classification de type discriminatif qui cherche à généraliser la régression logistique binaire au cas multiclasse.

On cherche à modéliser la probabilité conditionnelle de chaque classe C_l étant donnée le vecteur aléatoire $\mathbf{X}$. Cette probabilité conditionnelle est une fonction linéaire en $\mathbf{X}$.

$$\forall k = 1, ..., q : P(Y = C_k|X = x) = \frac{exp(a_{k0} + a_k^T x)}{\sum_{l=1}^{q} exp(a_{l0} + a_l^T x)}$$

Soit une matrice binaire $\mathbf{Y}$ de taille $(n \times q)$ qui permet de représenter l'appartenance d'un individu à une classe et de terme général :

$$y_{il} = \begin{cases} 1 & \text{si } x_i \in C_l \\ 0 & \text{sinon.} \end{cases}$$

En choisissant une **distribution multinomiale** pour modéliser la probabilité $P(Y|X)$ et en sous l'hypothèse **i.i.d** (variables indépendantes et identiquement distribuées) alors la vraisemblance s'écrit :

$$vr(\mathbb{P}) = \prod_{l=1}^{q} \prod_{i=1}^{n} P(Y = C_l|x_i; a_l)^{y_{il}}$$

Ce qui nous rappelle les **formules** de $P(w)$ ou $P(z)$ ou nous avions l'hypothèse d' i.i.d

$$P(w) = \prod_{j} P(w_j) = \prod_{jl} \beta_l^{w_{jl}} = \prod_{j=1}^{p} \prod_{l=1}^{L} P(w_{jl} = 1; \beta)^{w_{jl}} = \prod_{j=1}^{p} \prod_{l=1}^{L} P(w = l|j; \beta)^{w_{jl}}$$

$$P(z) = \prod_{i} P(z_i) = \prod_{ik} \alpha_k^{z_{ik}} = \prod_{i=1}^{n}\prod_{k=1}^{K} P(z_{ik} = 1; \alpha)^{z_{ik}} = \prod_{i=1}^{n}\prod_{k=1}^{K} P(z = k|i; \alpha)^{z_{ik}}$$

Cf : Cours d'Advanced Supervised Learning **J. Ah Pine**

6.4 1ere Idée d'implementation

Notre idée pour implémenter les dépendances temporelles et spatiales basée sur [15] est la suivante :

D'abord nous partons de l'équation de LBM

$$f(x; \theta) = \sum_{(z,w)\in Z*W} \prod_{ik} \alpha_k^{z_{ik}} \prod_{jl} \beta_l^{w_{jl}} \prod_{ijkl} f(x_{ij}, \theta_{kl})^{z_{ik}w_{jl}}$$

Essayons de définir $f(x_{ij}, \theta_{kl})$ de sorte à ce qu'il prenne en compte des dépendances en se basant sur l'article proposé par **cheam et al**.

$$f(x_{ij}, \theta_{kl}) = \prod_{i=1}^{n}\prod_{j=1}^{p}\sum_{n=1}^{N} \omega_{klijn}(\lambda_{kl}) \times \phi(c_{ij}|[M_t - M_{t-1}]'\beta_{kln} + c_{i(j-1)}, \sigma^2_{kln})$$

- c_{ij} : les coefficients de l'expansion de base après MFPCA
- N : Nombre de composants du modèle de mélange de régression polynomiale
- Q : Le degré du polynôme de régression
- $M_t = (1, m_t, ..., m_t^Q)$: Vecteur du Q-ieme degré de polynôme de
- $\beta_{kln} = (\beta_{kln0}, ..., \beta_{klnQ})$: Vecteur de coefficient du N-ieme composant
- $\phi(.|\mu, \sigma^2)$: densité de la distribution gaussienne multivariée avec μ : moyenne et σ^2 : variance

$\omega_{klijn}(\lambda_{kl})$ est le **poid** de ce modèle de mélange . Il dépend a la fois de des dimensions spatiales et temporelles

Vu que le **block** lui-même est maintenant défini comme un modèle de mélange , une nouvelle variable latente est nécessaire : $y_{ij} = (y_{ij1}, ..., y_{ijN})$

$$y_{ijn} = \begin{cases} 1 & \text{si } x_{ij} \in \text{n-ieme polynome de regression du kl-ieme composant} \\ 0 & \text{sinon.} \end{cases}$$

$$\omega_{gjtk}(\lambda_g) = \frac{\exp(\lambda_{gk1}u_j + \lambda_{gk2}v_j + \lambda_{gk3}m_t + \lambda_{gk4})}{\sum_{\ell=1}^{K} \exp(\lambda_{g\ell 1}u_j + \lambda_{g\ell 2}v_j + \lambda_{g\ell 3}m_t + \lambda_{g\ell 4})}$$
$$\forall\,(g,j,t,k), \tag{4}$$

avec $\lambda_{kl} = (\lambda_{kl11}, ..., \lambda_{klN4})$: les paramètres de la fonction de régression logistique.

Dans cette idée d'implémentation un **block** est lui-même défini comme étant une **modèle de mélange**.

6.5 2eme Idée d'implémentation

Dans cette deuxième idée basée sur [16], Nous modifions directement le $P(w)$ et/ou $P(z)$ de LBM en supposant ou non leurs indépendances

En partant de :

$$f(x;\theta) = \sum_{(z,w)\in Z*W} P(z,w) f(x|z,w;\theta)$$

Nous maintenons les hypothèses d'indépendance 1 et 4 :

- Clusters lignes et Clusters colonnes indépendants

- Pour une colonne et une ligne donnée , X_{ij} indépendants et identiquement distribués

Nous avons donc **3 possibilités** :

— Les Clusters **lignes indépendants** et identiquement distribués et les Clusters **colonnes dépendants** et identiquement distribués

— Les Clusters **colonnes indépendants** et identiquement distribués et les Clusters **lignes dépendants** et identiquement distribués

— Les clusters **lignes** et **colonnes dépendants**

Soit
$m = (m_1, ..., m_p)$: la grille temporelle de $\mathbf{X}$
$s = (s_1, ..., s_n)$: la grille spatiale de $\mathbf{X}$

$x_{ij} \in \mathbb{R}^s$ est la réalisation de X_{ij} correspondant au site i et au temps j.
Les coordonnées spatiales du site j sont données par le vecteur $s_i = (u_i, v_i)$

6.5.1 Les Clusters lignes i.i.d et les clusters colonnes dépendants (dépendances temporelles uniquement)

Comme vu plus haut , en ne supposant pas l'indépendance de la variable latente **w** , nous pouvons quand même modéliser $P(w)$ grâce a la **régression logistique multinomiale**.

$$P(w; \lambda_{jl}) = \prod_j P(w_j) = \prod_{jl} \frac{exp(\lambda_{l0} + \lambda_l^T m_j)}{\sum_{t=1}^{L} exp(\lambda_{t0} + \lambda_t^T m_j)}$$

$$f(x; \theta) = \sum_{(z,w) \in Z*W} \prod_{ik} \alpha_k^{z_{ik}} \prod_{jl} \frac{exp(\lambda_{l0} + \lambda_l^T m_j)}{\sum_{t=1}^{L} exp(\lambda_{t0} + \lambda_t^T m_j)} \prod_{ijkl} P(x_{ij}, \theta_{kl})^{z_{ik} w_{jl}}$$

L'inférence de ce nouveau modèle LBM se fait avec SEM-Gibbs(Stochastic Expectation-Maximisation)

- Expectation (Etape SE Modifiée)

1) Nous simulons $z^{(h+1)}|c, w^{(h)}$

$$p(z_{ik} = 1|c, w^{(h)}; \theta^{(h)}) = \frac{\alpha_k^{(h)} f_k(c_i|w^{(h)}; \theta^{(h)})}{\sum_{k'} \alpha_{k'}^{(h)} f_{k'}(c_i|w^{(h)}; \theta^{(h)})}$$

Avec $f_k(c_i|w^{(h)}; \theta^{(h)}) = \prod_{jl} p(c_{ij}, \theta_{kl}^{(h)})^{w_{jl}^{(h)}}$

2) Nous simulons maintenant $w^{(h+1)}|c, z^{(h+1)}$

$$p(w_{jl} = 1|c, z^{(h+1)}; \theta^{(h)}) = \frac{\pi_l(m; \lambda_{jl}^{(h)}) f_l(c_j|z^{(h+1)}; \theta^{(h)})}{\sum_{l'} \pi_{l'}(m; \lambda_{jl}^{(h)}) f_{l'}(c_j|z^{(h+1)}; \theta^{(h)})}$$

Avec $\pi_l(m; \lambda_{jl}^{(h)}) = \prod_j \frac{exp(\lambda_{l0} + \lambda_l^T m_j)}{\sum_{t=1}^{L} exp(\lambda_{t0} + \lambda_t^T m_j)}$ et $f_l(c_j|z^{(h+1)}; \theta^{(h)}) = \prod_{ik} p(c_{ij}, \theta_{kl}^{(h)})^{z_{ik}^{(h)}}$

- Maximisation (Etape M Modifiée)

Rien ne change si ce n'est que :

- $\beta_l^{(h+1)}$ n'est plus calculé par contre , nous calculons maintenant $\lambda_{jl}^{(h+1)}$

- $\lambda_{jl}^{(h+1)} = \arg \underset{\lambda_{jl}}{max} \sum_i \sum_j \sum_k \sum_l E[log(p(c, z^{(h+1)}, w^{(h+1)}|\theta)|\theta^{old})]\ log\pi_l(m; \lambda_{jl}^{(h)})$

$\lambda_{jl}^{(h+1)}$ => la solution de la regression logistique obtenu par l'algorithme de Newton-Raphson

6.5.2 Les Clusters colonnes i.i.d et les clusters lignes dependants (dependances spatiales uniquement)

Nous pouvons quand même modéliser $P(z)$ grace a la **regression logistique multinomiale**.

$$P(z; \lambda_{ik}) = \prod_i P(z_i) = \prod_{ik} \frac{exp(\lambda_{k0} + \lambda_{k1}^T u_i + \lambda_{k2}^T v_i)}{\sum_{s=1}^K exp(\lambda_{s0} + \lambda_{s1}^T u_i + \lambda_{s2}^T v_i)}$$

$$f(x; \theta) = \sum_{(z,w) \in Z*W} \prod_{ik} \frac{exp(\lambda_{k0} + \lambda_{k1}^T u_i + \lambda_{k2}^T v_i)}{\sum_{s=1}^K exp(\lambda_{s0} + \lambda_{s1}^T u_i + \lambda_{s2}^T v_i)} \prod_{jl} \beta_l^{w_{jl}} \prod_{ijkl} P(x_{ij}, \theta_{kl})^{z_{ik} w_{jl}}$$

L'inférence de ce nouveau modèle LBM se fait avec SEM-Gibbs(Stochastic Expectation-Maximisation)

- Expectation (Etape SE Modifiée)

1) Nous simulons $z^{(h+1)}|c, w^{(h)}$

$$p(z_{ik} = 1|c, w^{(h)}; \theta^{(h)}) = \frac{\pi_k(s; \lambda_{ik}^{(h)}) f_k(c_i|w^{(h)}; \theta^{(h)})}{\sum_{k'} \pi_{k'}(s; \lambda_{ik}^{(h)}) f_{k'}(c_i|w^{(h)}; \theta^{(h)})}$$

Avec $\pi_k(s; \lambda_{ik}^{(h)}) = \prod_i \frac{exp(\lambda_{k0} + \lambda_{k1}^T u_i + \lambda_{k2}^T v_i)}{\sum_{s=1}^K exp(\lambda_{s0} + \lambda_{s1}^T u_i + \lambda_{s2}^T v_i)}$ et $f_k(c_i|w^{(h)}; \theta^{(h)}) = \prod_{jl} p(c_{ij}, \theta_{kl}^{(h)})^{w_{jl}^{(h)}}$

2) Nous simulons maintenant $w^{(h+1)}|c, z^{(h+1)}$

$$p(w_{jl} = 1|c, z^{(h+1)}; \theta^{(h)}) = \frac{\beta_l^{(h)} f_l(c_j|z^{(h+1)}; \theta^{(h)})}{\sum_{l'} \beta_{l'}^{(h)} f_{l'}(c_j|z^{(h+1)}; \theta^{(h)})}$$

Avec $f_l(c_j|z^{(h+1)}; \theta^{(h)}) = \prod_{ik} p(c_{ij}, \theta_{kl}^{(h)})^{z_{ik}^{(h)}}$

- Maximisation (Etape M Modifiée)

Rien ne change si ce n'est que :

- $\alpha_k^{(h+1)}$ n'est plus calculé par contre , nous calculons maintenant $\lambda_{ik}^{(h+1)}$

- $\lambda_{ik}^{(h+1)} = \arg \underset{\lambda_{ik}}{max} \sum_i \sum_j \sum_k \sum_l E[log(p(c, z^{(h+1)}, w^{(h+1)}|\theta)|\theta^{old})]\ log\pi_k(s; \lambda_{ik}^{(h)})$

$\lambda_{ik}^{(h+1)} =>$ la solution de la régression logistique obtenu par l'algorithme de Newton-Raphson

6.5.3 Les Clusters colonnes et lignes dépendants (dépendances spatio-temporelle)

Nous pouvons quand même modéliser $P(z)$ et $P(w)$ grâce a la **régression logistique multinomiale**.

$$P(z; \lambda_{ik}) = \prod_i P(z_i) = \prod_{ik} \frac{exp(\lambda_{k0} + \lambda_{k1}^T u_i + \lambda_{k2}^T v_i)}{\sum_{s=1}^K exp(\lambda_{s0} + \lambda_{s1}^T u_i + \lambda_{s2}^T v_i)}$$

$$P(w; \lambda_{jl}) = \prod_j P(w_j) = \prod_{jl} \frac{exp(\lambda_{l0} + \lambda_l^T m_j)}{\sum_{t=1}^L exp(\lambda_{t0} + \lambda_t^T m_j)}$$

$$f(x; \theta) = \sum_{(z,w) \in Z*W} \prod_{ik} \frac{exp(\lambda_{k0} + \lambda_{k1}^T u_i + \lambda_{k2}^T v_i)}{\sum_{s=1}^K exp(\lambda_{s0} + \lambda_{s1}^T u_i + \lambda_{s2}^T v_i)} \prod_{jl} \frac{exp(\lambda_{l0} + \lambda_l^T m_j)}{\sum_{t=1}^L exp(\lambda_{t0} + \lambda_t^T m_j)} \prod_{ijkl} P(x_{ij}, \theta_{kl})^{z_{ik} w_{jl}}$$

L'inférence de ce nouveau modèle LBM se fait avec SEM-Gibbs(Stochastic Expectation-Maximisation)

- Expectation (Etape SE Modifiée)

1) Nous simulons $z^{(h+1)}|c, w^{(h)}$

$$p(z_{ik}=1|c,w^{(h)};\theta^{(h)}) = \frac{\pi_k(s;\lambda_{ik}^{(h)})f_k(c_i|w^{(h)};\theta^{(h)})}{\sum_{k'}\pi_{k'}(s;\lambda_{ik}^{(h)})f_{k'}(c_i|w^{(h)};\theta^{(h)})}$$

Avec $\pi_k(s;\lambda_{ik}^{(h)}) = \prod_i \frac{exp(\lambda_{k0}+\lambda_{k1}^T u_i+\lambda_{k2}^T v_i)}{\sum_{s=1}^K exp(\lambda_{s0}+\lambda_{s1}^T u_i+\lambda_{s2}^T v_i)}$ et $f_k(c_i|w^{(h)};\theta^{(h)}) = \prod_{jl} p(c_{ij},\theta_{kl}^{(h)})^{w_{jl}^{(h)}}$

2) Nous simulons maintenant $w^{(h+1)}|c,z^{(h+1)}$

$$p(w_{jl}=1|c,z^{(h+1)};\theta^{(h)}) = \frac{\pi_l(m;\lambda_{jl}^{(h)})f_l(c_j|z^{(h+1)};\theta^{(h)})}{\sum_{l'}\pi_{l'}(m;\lambda_{jl}^{(h)})f_{l'}(c_j|z^{(h+1)};\theta^{(h)})}$$

Avec $\pi_l(m;\lambda_{jl}^{(h)}) = \prod_j \frac{exp(\lambda_{l0}+\lambda_l^T m_j)}{\sum_{t=1}^L exp(\lambda_{t0}+\lambda_t^T m_j)}$ et $f_l(c_j|z^{(h+1)};\theta^{(h)}) = \prod_{ik} p(c_{ij},\theta_{kl}^{(h)})^{z_{ik}^{(h)}}$

- Maximisation (Étape M Modifiée)

Rien ne change si ce n'est que :

- $\alpha_k^{(h+1)}$ et $\beta_j^{(h+1)}$ ne sont plus calculés par contre , nous calculons maintenant $\lambda_{ik}^{(h+1)}$ et $\lambda_{jl}^{(h+1)}$

- $\lambda_{ik}^{(h+1)} = \arg\max_{\lambda_{ik}} \sum_i \sum_j \sum_k \sum_l E[log(p(c,z^{(h+1)},w^{(h+1)}|\theta)|\theta^{old})]\ log\pi_k(s;\lambda_{ik}^{(h)})$

- $\lambda_{jl}^{(h+1)} = \arg\max_{\lambda_{jl}} \sum_i \sum_j \sum_k \sum_l E[log(p(c,z^{(h+1)},w^{(h+1)}|\theta)|\theta^{old})]\ log\pi_l(m;\lambda_{jl}^{(h)})$

$\lambda_{ik}^{(h+1)}$ et $\lambda_{jl}^{(h+1)}$ => les solutions des regressions logistique obtenu par l'algorithme de Newton-Raphson

6.5.4 Calculs pour le cas de la dépendance spatiale uniquement

$$H(\theta|\theta^{old}) = E[log(p(c,z^{(h+1)},w^{(h+1)}|\theta)|\theta^{old})]$$
$$= \sum_{i,k} z_{ik}^{(h+1)} log(\alpha_k) + \sum_{j,l} log(\pi_l(m;\lambda_{jl}^{(h)})) + \sum_{i,j,k,l} z_{ik}^{(h+1)} w_{jl}^{(h+1)} log(p(c_{ij};\theta_{kl}))$$

Mise a jour de $\lambda_{jl}^{(h)}$

Pour obtenir $\lambda_{jl}^{(h+1)}$, il fait calculer $\frac{\partial H(\theta|\theta^{old})}{\partial \lambda_{jl}}) = \lambda_{0l}^{(h)} + < \lambda_l^{(h)}, m >_{\mathbb{R}^N}$

Ce qui revient à trouver les solutions de la régression logistique obtenu par l'algorithme de Newton-Raphson.

7 Implémentation et Résultats de notre solution

7.1 Implémentation

Un package R du nom de *funSpaTimeLBM* a été développé pour permettre de prendre en compte des dépendances spatiales, temporelles, spatio-temporelles .

Ce package "overwrite" le package **funLBM** et permet de faire du co-clustering avec classique avec LBM ou un co-clustering avec une prise en compte des dépendances.

7.1.1 Modification du package funLBM

La modification majeure par rapport au package funLBM réside dans le calcul des β_l et des α_k

Dans la solution initiale :

$$\alpha_k = \frac{\sum_i z_{ik}}{N}$$
$$\beta l = \frac{\sum_j w_{jl}}{P}$$

Dans la solution proposée en fonction des cas de figure :

$$\alpha_k = \frac{\sum_i solutions(glm(z_{ik} \sim U_i + V_i))}{N}$$
$$\beta l = \frac{\sum_j solutions(glm(w_{jl} \sim m_j))}{P}$$

7.1.2 Guide d'utilisation

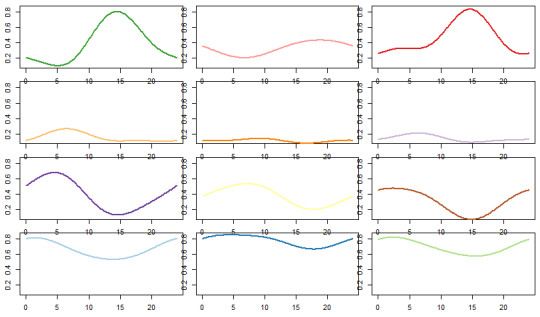
Une fois le package installé , Son utilisation est la suivante :

— **LBM classique** Ex : funSpaTimeLBM(list(data1,data2),K=4,L=3)

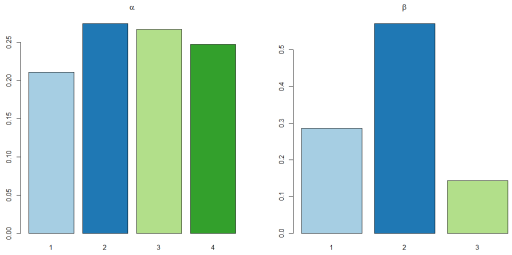
— **LBM avec dépendance temporelle** Ex : funSpaTimeLBM(list(data1,data2),K=4,L=3, timevar=c(1 :P))
— **LBM avec dépendance spatiale** Ex : funSpaTimeLBM(list(data1,data2),K=4,L=3, spatialvar=pos) avec pos = (U_i, V_i)
— **LBM avec dépendance spatio-temporelle** Ex : funSpaTimeLBM(list(data1,data2),K=4, L=3, spatialvar=pos, timevar=c(1 :P)) avec pos = (U_i, V_i)

7.2 Résultats

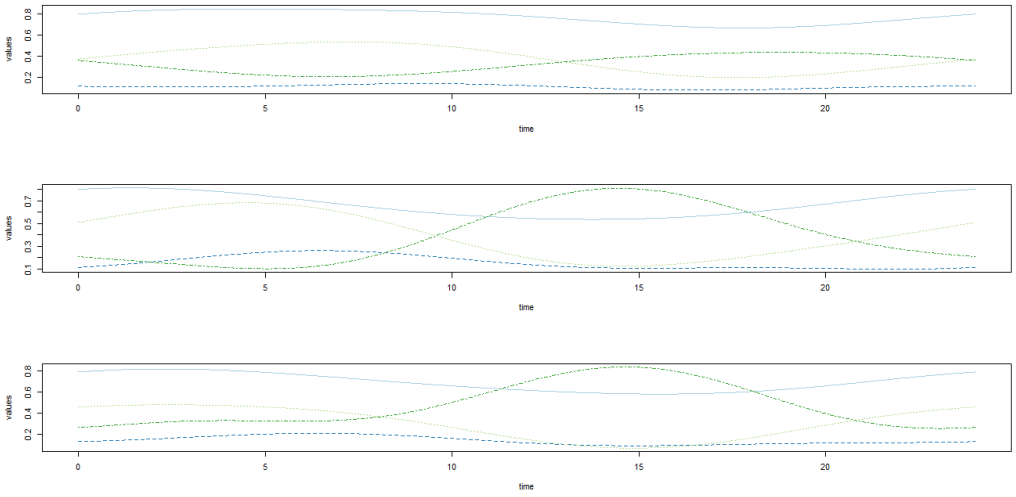
Le jeu de données qui a servi aux expériences est le dataset Velib provenant du système de partage de vélos de Paris, appelé Velib. Les données sont des profils de chargement des stations de vélos sur sept jours.
Les données ont été collectées toutes les heures pendant la période du dimanche 1er septembre au dimanche 7 septembre 2014.
C'est un jeu de données spatio-temporel car on dispose de la position (la longitude et la latitude des 1189 stations de vélos) pour 7 jours toutes les heures.
Cependant on ne dispose pas des labels des clusters lignes et colonnes.



(a) Blocks

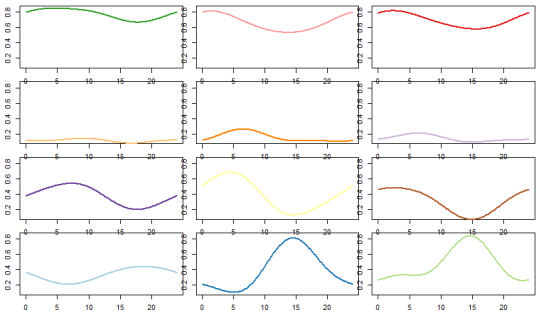


(b) Proportions

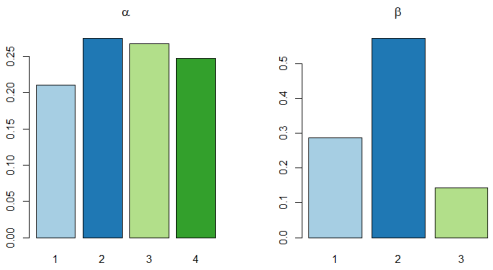


(c) Means

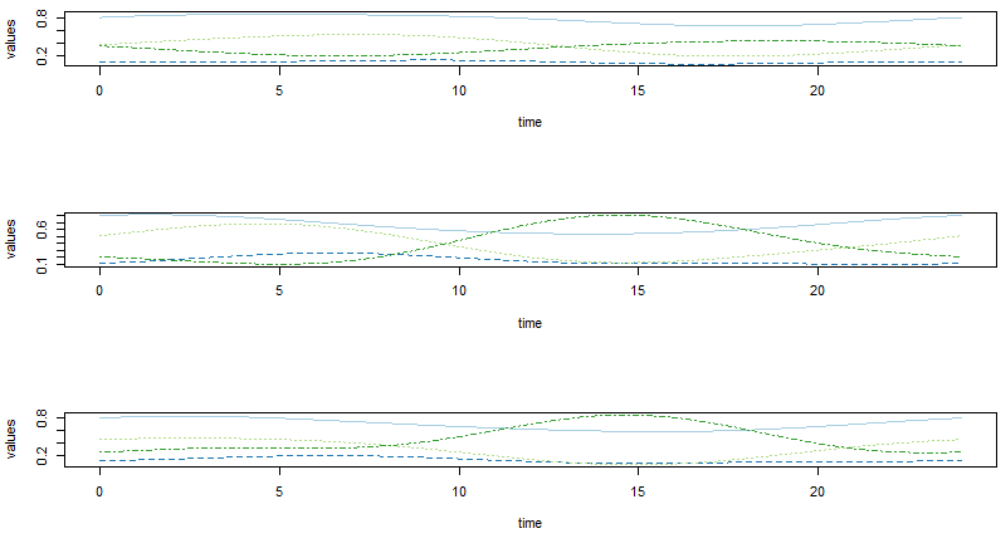
FIGURE 1 – LBM de base



(a) Blocks

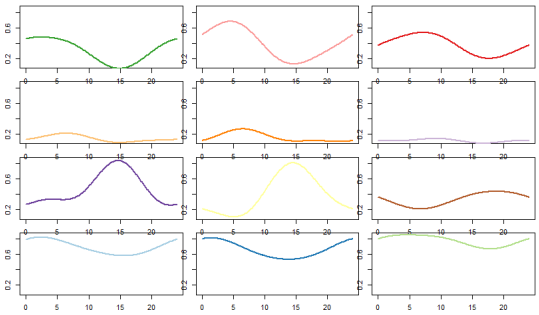


(b) Proportions

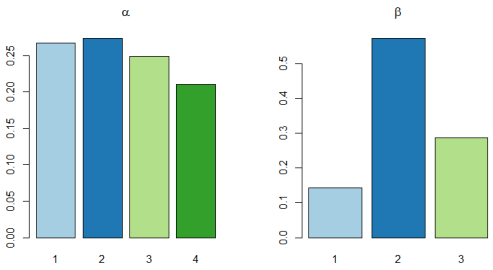


(c) Means

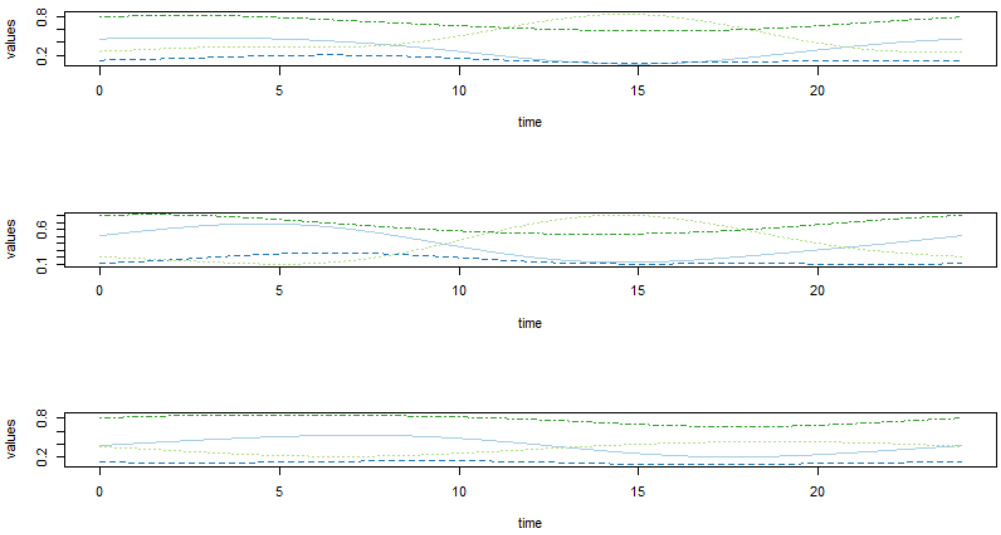
FIGURE 2 – Dépendance Temporelle



(a) Blocks

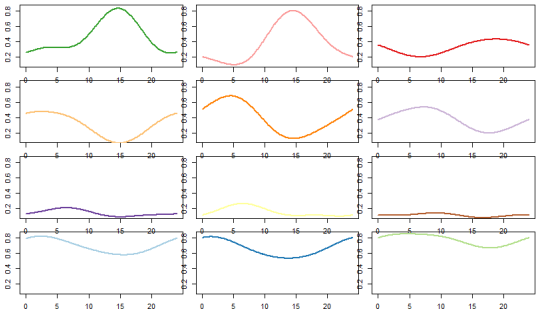


(b) Proportions

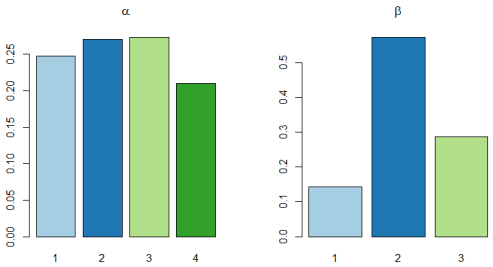


(c) Means

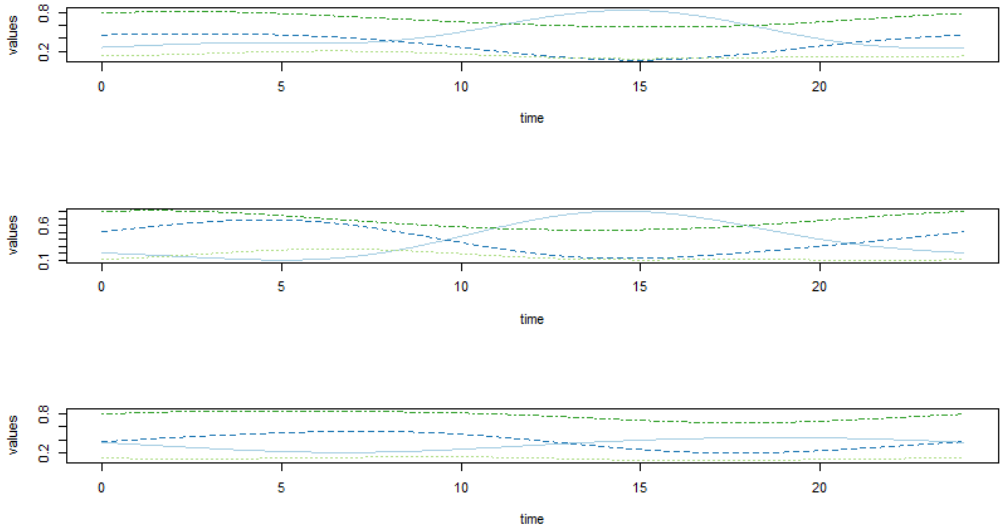
FIGURE 3 – Dépendance Spatiale



(a) Blocks



(b) Proportions



(c) Means

FIGURE 4 – Dépendance spatio-temporelle

Ne disposant pas de labels du dataset, nous ne pouvons pas utiliser la métrique **ARI** (Adjusted Rand Index) pour juger de la qualité du co-clustering. Mais tout de même nous remarquons qu'a chacune de nos expériences de sensibles **fluctuations** des résultats obtenus par rapport à la méthode **LBM** de base.

8 Conclusion

Dans les travaux proposés ci-dessus, Nous avons tenté de répondre a la problématique de la prise en compte de la dépendance spatiale ou temporelle dans un modèle de co-clustering nommé LBM.

En se basant sur les travaux de Cheam et al, et de Vandewalle et al , nous avons proposé d'utiliser la régression logistique binomiale pour essayer de prendre en compte ces dépendances.

Nous avons créer un package R contenant les modifications nécessaires pour mettre en évidence ces dépendances, puis nous avons testé ce code sur un dataset réel(Velib).

L'application du nouveau modèle implémenté n'a pas pu être mise en pratique sur des données simulées car nous n'avons pas réussi a simuler des dépendances dans ces jeux de données.

Il serait intéressant comme suite a ces travaux de valider la supériorité de ce modèle par rapport au modèle de base grâce notamment a un dataset simulé contenant des dépendances spatiales et/ou temporelles.

Références

[1] Charles Bouveyron, Julien Jacques, Amandine Schmutz, Fanny Simoes, Silvia Bottini, *Co-Clustering of Multivariate Functional Data for the Analysis of Air Pollution in the South of France*

[2] Hartigan J.A *Clustering Algorithms, Wiley New York, 1975 'Direct Clustering of data matrix ' Journal of the American Statistical Association, vol. 67 no. 337 pp.123-129, 197*

[3] Good I. *Categorisation of classification , Mathematics and Computer Sceince in Medecine and Biology, pp. 115-12, 1965*

[4] Fisher W. *Clustering and Aggregation in Economics John Hopkins Press , 1969*

[5] Tryon R., Bailyed *Cluster Analysis, McGraw-Hill, New York, 1970*

[6] Dubin R. Champoux J *Typology of empirical attributes : dissimilarity linkage analysis(DLA) Technical Report no. 3, University of California, 1970*

[7] Anderberg M.R *Cluster Analysis for Applications, Academic Press, New York, 1973.*

[8] Govaert G., *Algorithme de classification d'un tableau de contingence, 1st International Symposium on Data Analysis and Informatics, Versailles, INRIA, pp.487-500, 1977.*

[9] Govaert G. *Classification croisée, Thesis, University of Paris 6, France, 1983*

[10] Govaert G. and M. Nadif. *Co-Clustering. ISTE Ltd and John Wiley Sons Inc*

[11] Margot Selosse, Julien Jacques, Christophe Biernacki *Co-clustering de données textuelles et continues,*

[12] Amandine Schmutz, Julien Jacques, Charles Bouveyron, Laurence Chèze, Pauline Martin *Co-clustering de courbes fonctionnelles multivariées*

[13] Amandine Schmutz, Julien Jacques, Charles Bouveyron, Laurence Cheze, Pauline Martin *Données fonctionnelles multivariées issues d'objets connectés : une méthode pour classer les individus.*

[14] Gabriel Frisch, Jean-Benoist Leger, Yves Grandvalet *Learning from missing data with the Latent Block Model.*

[15] A. S. M. Cheam, M. Marbac, P. D. McNicholas *Model-based clustering for spatiotemporal data on air quality monitoring.*

[16] Vincent Vandewalle, Cristian Preda, Sophie Dabo *Clustering spatial functional data.*