

Propriétés émergentes du *multi-clustering* bayésien non paramétrique: Application aux données images multivues

Reda Khoufache*, Mohamed Djallel Dilmi*, Hanene Azzag*, Etienne Goffinet**
Mustapha Lebbah***

* Université Sorbonne Paris Nord, Villetaneuse, France

** Technology Innovation Institute, Abu Dhabi, United Arab Emirates

*** Université de Versailles - Université Paris Saclay, Versailles, France

1 Introduction

Dans le cas multivarié, le *clustering* infère uniquement une partition ligne, tandis que le *multi-clustering* infère une partition en colonne (partition de variables ou vues), et une partition ligne pour chaque vue. La modélisation bayésienne non paramétrique permet d'estimer le nombre de composantes durant l'inférence en mettant une distribution a priori sur les paramètres du modèle.

Dans Mansinghka et al. (2009), les auteurs ont introduit un modèle de catégorisation croisée, et Guan et al. (2010) ont proposé un modèle de *multi-clustering* bayésien non paramétrique. Ces deux travaux partagent la même définition du modèle, qui met d'abord un a priori sur la partition colonne, qui estime automatiquement le nombre de *clusters* colonnes, ensuite met un a priori indépendant sur les proportions de chaque partition ligne.

1.1 Définition du modèle

Notons $X \in \mathbb{R}^{n \times p \times d}$ l'espace latent obtenu après une certaine transformation du jeu de données. Soit H le nombre de *clusters* de variables, v la partition de variables, Z une matrice indicatrice $n \times H$, des partitions lignes. Le modèle est défini comme suit :

$$\begin{aligned} x_{i,j} \mid \{v_j = h, z_i^h = k, \theta_k^h\} &\sim \mathcal{N}(\theta_k^h), \\ \theta_k^h &\sim G_0, v_j \sim \text{Mult}(\eta), z_i^h \sim \text{Mult}(\pi_h), \\ \eta_j(\mathbf{r}) &= r_j \prod_{j'=1}^{j-1} (1 - r_{j'}), r_j \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(1, \gamma), \\ \pi_j^h(\mathbf{t}^h) &= t_j^h \prod_{j'=1}^{j-1} (1 - t_{j'}^h), t_j^h \stackrel{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha_h), \\ \gamma &\sim \text{Gamma}(a_\gamma, b_\gamma), \alpha_h \sim \text{Gamma}(a_\alpha, b_\alpha). \end{aligned}$$

Où les proportions de la partition de variables η et des partitions lignes π_h suivent le processus *Stick-Breaking* Sethuraman (1994).