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 nomble 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 :

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

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

2 Le framework proposé

Le *framework* que nous proposons est constituée de deux étapes : La première étape consiste à extraire des caractéristiques des images et de réduire la dimension de l'espace de représentation. Le jeu de données d'images est réarrangé sous forme d'un tableau bi-dimensionnel où les lignes représentent les observations, les colonnes sont les différentes variables. Dans la première étape, nous proposons d'utiliser un *Vision Transformers* pré-entraîné comme extracteur de caractéristiques. En raison du fléau de la dimension et de la complexité algorithmique, une analyse en composantes principales (*PCA*) est appliquée afin de réduire la dimension de l'espace de représentation. La seconde étape réalise le *multi-clustering* bayésien non paramétrique qui estime automatiquement le nombre de blocs. Le *framework* complet est illustré dans la figure 1.

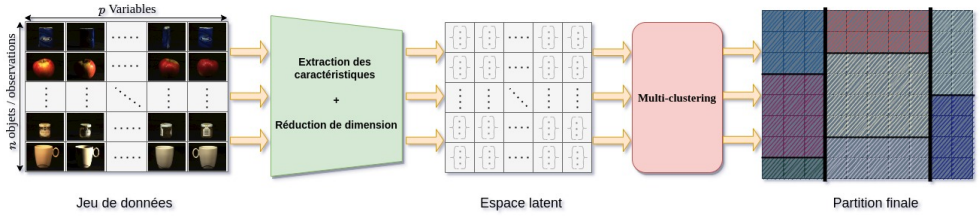

FIG. 1 – Le *framework* complet.

3 Conclusion

Dans ce papier, nous avons proposé un *framework* pour inférer de multiples solutions de *clustering* sur un jeu de données d'images multivues. Notre *framework* combine un *vision transformers* pré-entraîné avec un algorithme du *multi-clustering*. Nous avons proposé une approche bayésienne non paramétrique, qui permet d'estimer la structure du modèle durant l'inférence. Notre approche a fournit des blocs homogènes et cohérents.

Références

Guan, Y., J. Dy, D. Niu, et Z. Ghahramani (2010). Variational inference for nonparametric multiple clustering.

Mansinghka, V., E. Jonas, C. Petschulat, B. Cronin, P. Shafto, et J. Tenenbaum (2009). Cross-categorization : A method for discovering multiple overlapping clusterings.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica Sinica 4*(2), 639–650.