# Co-clustering de données mixtes à base des modèles de mélange

Aichetou Bouchareb*, Marc Boullé*, Fabrice Rossi**

*Orange Labs
prenom.nom@orange.com
**SAMM EA 4534 - Université Paris 1 Panthéon-Sorbonne
prenom.nom@univ-paris1.fr

**Résumé.** La classification croisée (co-clustering) est une technique non supervisée qui permet d'extraire la structure sous-jacente existante entre les lignes et les colonnes d'une table de données sous forme de blocs. Plusieurs approches ont été étudiées et ont démontré leur capacité à extraire ce type de structure dans une table de données continues, binaires ou de contingence. Cependant, peu de travaux ont traité le co-clustering des tables de données mixtes. Dans cet article, nous étendons l'utilisation du co-clustering par modèles à blocs latents au cas des données mixtes (variables continues et variables binaires). Nous évaluons l'efficacité de cette extension sur des données simulées et nous discutons ses limites potentielles.

## 1 Introduction

La classification croisée a pour objectif de réaliser une classification jointe des lignes et des colonnes d'un tableau de données. Proposée par Good (1965) puis par Hartigan (1975), la classification croisée est une extension de la classification simple (clustering) qui permet d'extraire la structure sous-jacente dans les données sous forme de groupes de lignes et groupes de colonnes. L'avantage de cette technique, par rapport à la classification simple, réside dans l'étude *simultanée (jointe)* des lignes et des colonnes qui permet d'extraire un maximum d'informations sur la dépendance entre elles. L'utilité du co-clustering réside dans sa capacité de créer des groupes facilement interprétables et dans sa capacité de réduction d'une grande table de données en une matrice significativement plus petite et ayant la même structure que les données originales. Le traitement de la matrice résumée permet d'étudier et de prendre des décisions sur les données originales tout en réduisant significativement les coûts de calcul en temps et en mémoire.

Depuis son introduction, plusieurs méthodes ont été développées pour effectuer une classification croisée (Bock (1979); Cheng et Church (2000); Dhillon et al. (2003); Xu et al. (2010)). Ces méthodes diffèrent principalement dans le type des données étudiées (continues, binaires ou de contingence), les hypothèses considérées, la méthode d'extraction utilisée et la forme souhaitée pour les résultats (classification stricte ou floue, hiérarchie, etc.). L'une des approches les plus connues est celle de la classification par modèles à blocs latents qui est basée sur des