Subspace clustering sur données incomplètes par imputation multiple

Yasmine Agliz, Vincent Audigier, Mohamed Nadif, Ndèye Niang

Subspace clustering sur données incomplètes par imputation multiple

Yasmine AGLIZ[*,**], Vincent AUDIGIER[**]
Mohamed NADIF[***], Ndèye NIANG [**]

[*]51 Rue de Lille,75007
yasmine.agliz@gmail.com,
[**]2 Rue conté, 75003
vincent.audigier@cnam.fr
ndeye.niang_keita@cnam.fr
[***]45 rue des saint pères,75006
mohamed.nadif@u-paris.fr

Résumé. Nous nous intéressons à la classification non supervisée en grande dimension en présence d'observations incomplètes. Nous proposons une approche d'imputation multiple avec comme modèle d'analyse la méthode Reduced KMeans de *subspace clustering*. L'agrégation des résultats, extension des règles de Rubin, repose d'une part sur des méthodes de consensus de partitions et d'autre part sur la recherche d'un sous-espace commun de représentation obtenu à travers l'Analyse factorielle multiple. Une étude de simulation montre de bonne performances d'une part en termes de classification à travers l'indice de Rand ajusté (ARI) et la Normalized Mutual Information (NMI) et d'autre part en terme de sous-espace réduit à travers le coefficient RV.

1 Introduction

Dans un cadre de grande dimension, les classes sont souvent décrites par des sous-espaces de variables et donc de nombreuses variables deviennent non pertinentes pour la tâche de classification. Ces variables peuvent alors pénaliser l'apprentissage des algorithmes classiques de classification en masquant les classes ; il devient ainsi difficile pour ces algorithmes de retourner la bonne partition (Parsons et al., 2004; Yamamoto et Hwang, 2014). Pour pallier à cette inefficacité, les algorithmes de *Subspace Clustering* ont notamment été proposés. L'objectif de ces derniers est de retrouver des classes et leur sous-espaces caractéristiques. Ces derniers peuvent être obtenus à travers des méthodes de sélection de variables basées sur des systèmes de pondération (Diallo et al., 2021) ou par combinaisons linéaires de l'ensemble des variables. Parmi ces méthodes combinaisons linéaires des variables, on peut citer une l'*approche tandem* (Hubert et Arabie, 1985), consistant à appliquer la classification sur les premières composantes principales d'une analyse factorielle. Toutefois, cette approche reste critiquable car si les composantes principales maximisent l'inertie de projection du nuage de points, rien ne garantit qu'elle maximise la dispersion des centres de gravité des classes recherchées (De Soete et Carroll, 1994). Une autre approche consiste à rechercher simultanément la partition des individus

et les composantes optimales pour la tâche de classification comme proposé dans les méthodes Factorial K-means (FKM) et Reduced K-means (RKM) (De Soete et Carroll, 1994; Timmerman et al., 2010).

Par ailleurs, le grand nombre de variables rend incontournable le problème des données manquantes, constituant ainsi une difficulté supplémentaire. La tâche de classification par les approches géométriques dans le cadre de données incomplètes a cependant fait l'objet de différents travaux, notamment en "petite dimension". Une première approche, généralement peu recommandée, consiste à se ramener à un jeu de données complet en supprimant les observations ou variables incomplètes. D'autres méthodes, dites "directes", plus performantes, consistent à s'accommoder des données manquantes. Ces méthodes reviennent à adapter le critère d'optimisation de telle sorte qu'il ne se base que sur les valeurs observées. Plusieurs méthodes de clustering connues on été adaptées aux données incomplètes telles que k-means (Chi et al., 2016) et fuzzy c-means (Hathaway et Bezdek, 2001).

On retrouve aussi dans la littérature des approches dites d'imputation basées sur le remplacement des valeurs manquantes. On distingue les méthodes d'imputation simple qui consiste à remplacer une fois la valeur manquante et les méthodes d'imputation multiple permettant de prendre en compte l'incertitude liée à l'imputation contrairement au cas simple.

Nous nous intéressons aux méthodes de subspace clustering sur données incomplètes. Plus précisément, nous proposons d'utiliser l'imputation multiple pour la gestion des données manquantes dans le cadre de la méthode Reduced K-means (RKM). Cette dernière permet d'obtenir une partition des individus en classes ainsi que le sous-espace qui lui est associé. Dans le cas de l'imputation multiple, cela conduit à un ensemble de partitions et plusieurs sous-espaces associés qu'il faut ensuite agréger.

Les travaux de Audigier et Niang (2022), proposent une méthode permettant d'obtenir un consensus des partitions obtenues. Dans ce travail, nous nous intéressons à l'étape d'agrégation des différents sous-espaces obtenus. Cette étape peut être complexe impliquant de combiner des structures de données potentiellement différentes, car issues de diverses imputations, tout en préservant la cohérence et la pertinence des classes formées. Les travaux de Josse et Husson (2012) utilisent ces variations pour évaluer la stabilité des paramètres d'une analyse factorielle à des fins exploratoires. Nous proposons d'utiliser l'analyse factorielle multiple (AFM) (Pages, 2004) pour trouver un espace compromis représentant l'agrégation des sous-espaces.

Dans la section suivante nous présentons la méthode RKM. La section 3 est dédiée à la classification de données incomplètes par imputation multiple. La méthode proposée est présentée en section 4 en précisant d'abord les différentes étapes puis en présentant son évaluation via une étude par simulation.

2 Reduced K-means

Reduced K-means peut-être vue comme une adaptation de la populaire méthode des K-means dans un contexte de *subspace clustering*, avec le critère suivant :

$$RKM(\mathbf{A}, \mathbf{F}, \mathbf{U}|c, q) = \min_{\mathbf{U},\mathbf{F},\mathbf{A}} ||\mathbf{X} - \mathbf{U}\mathbf{F}\mathbf{A}^T||_F^2 \quad (1)$$

où $\mathbf{X}$ est la matrice des données de dimensions $(I \times J)$ comportant en ligne les individus et en colonne les variables, $\mathbf{U}$ la matrice d'appartenance des observations aux c classes, composée de 0 et 1, de dimensions $(I \times c)$, et $||.||_F$ la norme de Frobenius.

Reduced K-means consiste à reformuler le critère du kmeans de façon à ce que la matrice des centroïdes s'exprime sous la forme d'un produit matriciel $\mathbf{FA}^\top$ où $\mathbf{F}$ de dimensions ($c \times q$) est la matrice des centroïdes dans un espace réduit de dimension q et $\mathbf{A}$ de dimensions ($J \times q$) est une matrice de loadings déterminant la contribution de chaque variable à la structure en groupe des observations. Ceci revient à minimiser la somme des distances au carré entre les observations et les centroïdes de l'espace réduit, reconstitués dans l'espace initial. D'un point de vue modélisation, la méthode peut aussi se présenter selon

$$\mathbf{X} = \mathbf{UFA}^T + \mathbf{E} \tag{2}$$

avec $\mathbf{E}$ une matrice ($I \times J$) de résidus indépendants identiquement distribués selon une loi normale centrée.

L'optimisation du critère (1) s'effectue en alternant entre la recherche de la partition, via la matrice $\mathbf{U}$, obtenue par K-means, et la mise à jour du sous-espace, via les matrices $\mathbf{F}$ et $\mathbf{A}$, obtenues par décomposition en valeurs singulières (Terada, 2014). Afin d'éviter la convergence vers un minimum local, plusieurs initialisations sont nécessaires.

3 Classification sur données incomplètes

Nous nous plaçons dans le cadre classique où les données sont manquantes au hasard (Rubin, 1976), également appelées données *Missing At Random* (MAR). L'imputation multiple se déroule en trois étapes bien définies :

1. Imputer plusieurs fois le jeu de données selon un modèle dit d'imputation
2. Appliquer un modèle d'analyse sur chaque tableau
3. Agréger les résultats selon les règles de Rubin (Rubin, 1976).

Ces règles consistent à agréger les paramètres du modèle d'analyse et à quantifier la variabilité de l'imputation (Basagaña et al., 2013; Bruckers et al., 2017).

Dans le cadre de la classification, Audigier et Niang (2022) ont adapté cette méthodologie en imputant les données selon un modèle joint (JM), une méthode d'imputation qui prend en compte la structure en groupe des données (Kim et al., 2014; Van Buuren, 2018). Ensuite, chaque jeu de données imputé est classifié par un algorithme dédié et les partitions obtenues sont agrégées à travers des méthodes de consensus à l'aide de la méthode Non-negative Matrix Factorization (NMF).

4 Méthode proposée

L'imputation multiple a été abordée dans un contexte de clustering avec une attention particulière accordée au consensus de partitions. Le contexte du subspace clustering mène à considérer le consensus des sous-espaces. Les méthodes factorielles d'analyse de tableaux multiples, telles que l'AFM, permettent en particulier de trouver un sous-espace compromis. Cet espace commun pour la représentation des individus décrits par l'ensemble des tableaux est obtenu en prenant en compte leur structuration à travers une pondération spécifique à chaque tableau.

4.1 Principales étapes de la méthode

1. Imputer M fois le jeu de données avec un modèle joint (JM)
2. Classifier chacun des M jeux de données imputées par Reduced K-means .
3. Agréger les M partitions avec la méthode Non-negative Matrix Factorization (NMF).
4. Agréger les M différents sous-espaces par la méthode AFM afin d'obtenir un seul sous-espace final.

4.2 Évaluation

Afin d'évaluer la méthode proposée, nous nous sommes inspirés du plan de simulation de Terada (2014). Ceci implique de définir les trois matrices $\mathbf{U}$, $\mathbf{F}$, $\mathbf{A}$ et $\mathbf{E}_{\text{sim}}$. La matrice $\mathbf{X}$ est constituée de la sorte :

$$\mathbf{X} = \mathbf{UFA}^T + \mathbf{E}_{\text{sim}} \tag{3}$$

Nous considérons une matrice $\mathbf{X}$ de dimensions $(I \times J)$, avec p_1 le nombre des variables informatives, p_2 le nombre des variables de bruit corrélées entres elles, et p_3 le nombre des variables de bruit indépendantes $(J = p1 + p2 + p3)$. Dans ce plan de simulation, la matrice $\mathbf{U}$ est générée grâce à une loi multinomiale avec des probabilités égales. La matrice de centroïdes $\mathbf{F}$ est générée à partir d'une distribution uniforme q-dimensionnelle sur $[-O, O]^q$. Avec O un paramètre qui permets de gérer la séparabilité des classes. Les matrices $\mathbf{A}$ et Σ_J sont ensuite construites de la sorte :

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^{*T} & \mathbf{0}^T_{q\times(p2+p3)} \end{bmatrix} \quad \text{et} \quad \Sigma_J = \begin{bmatrix} I_{p1} & \mathbf{0}_{p1\times p2} & \mathbf{0}_{p1\times p3} \\ \mathbf{0}_{p2\times p1} & \Sigma_{p2} & \mathbf{0}_{p2\times p3} \\ \mathbf{0}_{p3\times p1} & \mathbf{0}_{p3\times p2} & I_{p3} \end{bmatrix}$$

avec $\mathbf{A}^*$ une matrice orthogonale de dimension $p1 \times q$ générée de manière aléatoire. Chaque élément de $\mathbf{E}_{\text{sim}}$, est généré à partir de la distribution normale J-dimensionnelle $\mathcal{N}(\mathbf{0}, \Sigma_J)$. Avec Σ_{p2} la sous-matrice de dimensions $(p2 \times p2)$ de terme σ_{ij} $(1 \leq i, j \leq p_2)$ valant 1 pour $i = j$ et 0.25 sinon.

Pour un scénario, 100 jeux de données sont générés en se basant sur des matrices fixes $\mathbf{U}$, $\mathbf{F}$ et $\mathbf{A}$, en faisant varier $\mathbf{E}_{\text{sim}}$. Nous faisons varier les paramètres du plan de simulation pour obtenir différents scénarios, tels que des classes qui se chevauchent et des classes déséquilibrées. Pour chaque jeu de données, des données manquantes sont générées selon un mécanisme MCAR (Missing Completely at Random) et MAR (Missing at Random) pour différents taux de données manquantes (25% et 40%).

Les performances de la méthode sont mesurées selon l'indice de Rand ajusté (ARI) et la Normalized Mutual Information (NMI) entre les différentes partitions obtenues (M partitions et la partition consensus) et la véritable partition des jeux de données. Ensuite, le coefficient RV est utilisé pour comparer les matrices des composantes principales associées aux M sous espaces obtenues, l'espace compromis et la matrice de composantes principales utilisées pour générer les données.

Nous comparons notre approche d'imputation multiple à une approche dite directe, qui consiste à appliquer une version du Reduced K-means qui s'accommode aux données manquantes Agliz et al. (2024) (Reduced KPOD). Pour cette méthode, nous utilisons les mêmes paramètres (c et q) et une première initialisation des données manquantes par KNN.

Ci-dessous un exemple des premiers résultats obtenus pour un jeu de données d'un scénario : I=1000, J=50 (p3=p2=p1-1), c=10, q=5, O=50 ; Jeu de données n°1 ; 25% de données manquantes (MCAR) ; M=20

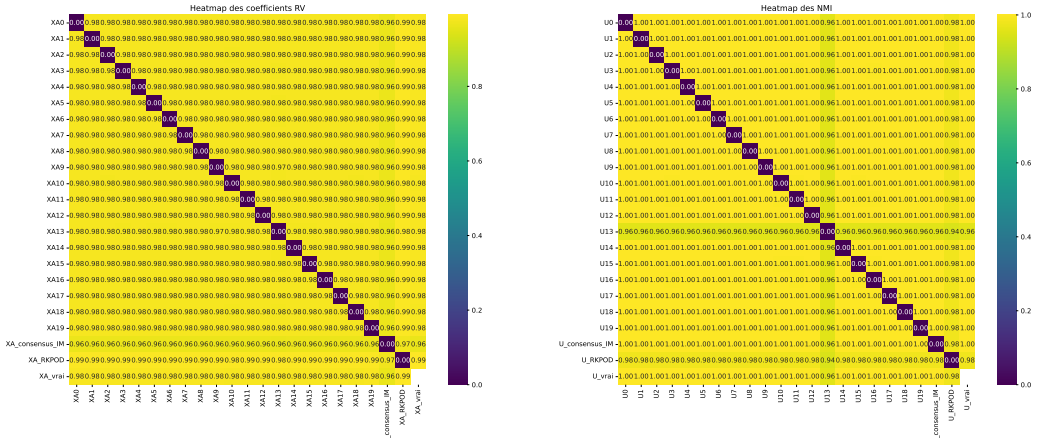

FIG. 1 – *Scores d'évaluation entre les matrices d'imputation multiple ($M = 20$), leurs résultats de consensus et les résultats du Reduced KPOD. Cette figure montre les coefficients RV (à gauche), et les indices de NMI (à droite) entre les étiquettes dérivées des matrices U.*

Sur ce premier jeu de données (Figure 1), on observe que le score NMI obtenu par imputation multiple est supérieur à celui du Reduced KPOD ; nous observons la même tendance pour le score ARI. Cependant, le consensus des composantes principales réalisé par l'AFM montre des résultats légèrement inférieurs, bien que toujours très satisfaisants.

5 Conclusion

Nous avons présenté une approche basée sur la méthode RKM pour le *subspace clustering* en présence de données incomplètes. Elle fournit une partition des observations ainsi que le sous-espace de représentation. La méthode appliquée à des données simulées donne des résultats satisfaisants en terme d'indice de qualité de la classification et d'identification du sous-espace pertinent pour les classes. D'autres travaux sont en cours pour une évaluation plus approfondie des performances, notamment à travers d'autres jeux de données.

Références

Agliz, Y., V. Audigier, et N. Niang (2024). Subspace clustering sur données incomplètes. In *Les 55èmes Journées de Statistique*.

Audigier, V. et N. Niang (2022). Clustering with missing data : which equivalent for rubin's rules ? *Advances in Data Analysis and Classification*, 1–35.

Basagaña, X., J. Barrera-Gómez, M. Benet, J. M. Antó, et J. Garcia-Aymerich (2013). A framework for multiple imputation in cluster analysis. *American journal of epidemiology 177*(7), 718–725.

Bruckers, L., G. Molenberghs, et P. Dendale (2017). Clustering multiply imputed multivariate high-dimensional longitudinal profiles. *Biometrical Journal 59*(5), 998–1015.

Chi, J. T., E. C. Chi, et R. G. Baraniuk (2016). k-pod : A method for k-means clustering of missing data. *The American Statistician 70*(1), 91–99.

De Soete, G. et J. D. Carroll (1994). K-means clustering in a low-dimensional euclidean space. In *New approaches in classification and data analysis*, pp. 212–219. Springer.

Diallo, A. W., N. Niang, et M. Ouattara (2021). Sparse subspace k-means. In *2021 International Conference on Data Mining Workshops (ICDMW)*, pp. 678–685. IEEE.

Hathaway, R. J. et J. C. Bezdek (2001). Fuzzy c-means clustering of incomplete data. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 31*(5), 735–744.

Hubert, L. et P. Arabie (1985). Comparing partitions. *Journal of classification 2*, 193–218.

Josse, J. et F. Husson (2012). Handling missing values in exploratory multivariate data analysis methods. *Journal de la société française de statistique 153*(2), 79–99.

Kim, H. J., J. P. Reiter, Q. Wang, L. H. Cox, et A. F. Karr (2014). Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics 32*(3), 375–386.

Pages, J. (2004). Multiple factor analysis : Main features and application to sensory data. *Revista Colombiana de Estadística 27*(1), 1–26.

Parsons, L., E. Haque, et H. Liu (2004). Subspace clustering for high dimensional data : a review. *SIGKDD Explor. 6*, 90–105.

Rubin, D. B. (1976). Inference and missing data. *Biometrika 63*(3), 581–592.

Terada, Y. (2014). Strong consistency of reduced k-means clustering. *Scandinavian Journal of Statistics 41*(4), 913–931.

Timmerman, M. E., E. Ceulemans, H. A. Kiers, et M. Vichi (2010). Factorial and reduced k-means reconsidered. *Computational Statistics & Data Analysis 54*(7), 1858–1871.

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC press.

Yamamoto, M. et H. Hwang (2014). A general formulation of cluster analysis with dimension reduction and subspace separation. *Behaviormetrika 41*(1), 115–129.

Summary

We are interested in high-dimensional clustering in the presence of incomplete observations. Here we propose a multiple imputation approach using a Reduced KMeans algorithm as the analysis model. The aggregation of the results, which is an extension of Rubin's rules, relies on both consensus partition methods and the search for a common subspace obtained through Multiple Factor Analysis. A simulation study shows good performance in terms of clustering using the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI), as well as in terms of the subspace using the RV coefficient.