Subspace clustering sur données incomplètes

Yasmine Agliz [1,2] & Vincent Audigier [3] & Ndèye Niang [4]

[1] *Laboratoire CEDRIC, équipe MSDMA, CNAM, Fance, yasmine.agliz@caissedesdepots.fr*
[2] *Caisse des dépôts, DAJCD, France*
[3] *Laboratoire CEDRIC, équipe MSDMA, CNAM, Fance, vincent.audigier@cnam.fr*
[4] *Laboratoire CEDRIC, équipe MSDMA, CNAM, Fance, ndeye.niang_keita@cnam.fr*

Résumé. Nous proposons ici une nouvelle approche pour la classification non-supervisée sur données incomplètes dans le cadre d'un grand nombre de variables. Cette approche intitulée *Reduced K-pod* s'appuie sur la formulation d'un critère de Reduced K-means calculable sur des données incomplètes sur le modèle de la méthode *K-pod*. Un algorithme d'optimisation du critère est proposé et sa convergence monotone est garantie. Cette méthode est ensuite évaluée par une étude par simulation mettant en évidence l'apport de la méthode par rapport aux approches géométriques concurrentes gérant soit les données manquantes (*K-pod*), soit les données manquantes et la grande dimension (par ACP itérative suivie de K-means). Les premiers résultats obtenus indiquent de meilleures performances en termes d'indice de Rand Ajusté mettant ainsi en évidence l'intérêt de l'approche pour la gestion de la grande dimension et des données manquantes en classification.

Mots-clés. Subspace clustering, classification, données grande dimension, données manquantes au hasard.

Abstract. A new approach for clustering on incomplete data within the framework of a large number of variables is proposed. This approach, entitled *Reduced K-pod*, is based on the formulation of a Reduced K-means criterion computable on incomplete data, in lines with the *K-pod* method. An algorithm for optimising the criterion is proposed with monotonic convergence ensured. This method is then evaluated using a simulation study highlighting the properties of the method compared with competitive geometric clustering approaches handling either missing data (*K-pod*) or missing data and high dimensionality (by iterative PCA followed by K-means). The preliminary obtained results indicate better performances in terms of Adjusted Rand Index, highlighting the interest of the approach for addressing high dimensionality and missing data in clustering.

Keywords. Subspace clustering, clustering, high dimensional data, Missing At Random data.

1 Introduction

Nous nous intéressons au problème de la classification automatique d'individus décrits par un grand ensemble de variables, ceci en nous focalisant sur approches géométriques. Dans ce cadre de grande dimension, de nombreuses variables deviennent non pertinentes pour la tâche de classification et les classes sont alors souvent décrites par des sous-espaces de variables. Or, ces variables non-pertinentes peuvent pénaliser l'apprentissage des algorithmes classiques de classification [1, 2].

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 [3] ou par combinaisons linéaires de l'ensemble des variables. Parmi ces dernières, on peut citer l'*approche tandem* [4], consistant à appliquer la classification sur les premières composantes principales d'une analyse factorielle. Toutefois, cette approche reste critiquable car si les composantes principales maximise l'inertie de projection du nuage de points, rien ne garantit qu'elle maximise la dispersion des centres de gravité des classes recherchées [5, 6]. Une autre approche consiste alors à rechercher simultanément la partition des individus et les composantes optimales pour la tâche de classification comme proposé dans les méthodes Reduced K-means [5] et Factorial K-means [6] (en abrégé, *RKM* et *FKM* respectivement).

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, ou en remplaçant les valeurs manquantes par imputation simple. D'autres approches, plus performantes, basées sur l'imputation multiple [7] ont également été proposées. On retrouve aussi dans la littérature des approches consistant à adapter les critères des méthodes de classification au caractère incomplet des données, comme proposé dans le cadre du k-means via l'approche *K-pod* [8], ou dans le cadre du fuzzy c-means et du k-prototype [9, 10].

Dans ce travail, nous nous intéresserons précisément à la classification sur données incomplètes dans le cadre d'un grand nombre de variables. Pour cela, nous définissons d'une part un nouveau critère, similaire à celui optimisé dans la méthode Reduced K-means, mais calculable sur données incomplètes, et d'autre part, un algorithme d'optimisation convergent de façon monotone, inspiré de celui de l'approche *K-pod*. On se place dans le cadre classique où les données sont manquantes au hasard [11] aussi appelées données *(Missing At Random)* (MAR).

La deuxième section présente les méthodes RKM et FKM, tandis que la troisième section présente la méthode *K-pod*. La méthode proposée, dénommée Reduced *K-pod*, fait l'objet de la quatrième section. Enfin, celle-ci est évaluée en cinquième section via une étude par simulation.

2 *Subspace clustering* par RKM et FKM

Ces deux approches peuvent être vues comme des adaptations de la populaire méthode des K-means dans un contexte de *subspace clustering*. La méthode des K-means peut en effet se présenter comme le problème de minimisation du critère suivant :

$$KM(\mathbf{C}, \mathbf{U}|c) = \min_{\mathbf{C},\mathbf{U}} ||\mathbf{X} - \mathbf{UC}||_F^2 \tag{1}$$

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 $(I \times c)$, $\mathbf{C}$ la matrice de centroïdes dans l'espace initial de dimensions $(c \times J)$, et $||.||_F$ la norme de Frobenius.

La méthode RKM consiste alors à reformuler ce critère de façon à ce que la matrice $\mathbf{C}$ 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. On obtient alors le critère suivant :

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

D'un point de vue modélisation, la méthode peut aussi se présenter selon

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

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

Là où le Reduced K-means vise à minimiser la somme des distances au carré entre les observations et les centroïdes de l'espace réduit, reconstitués dans l'espace initial, le Factoriel K-means lui consiste à minimiser la somme des distances au carré entre les observations dans l'espace réduit et leurs centroïdes correspondants dans ce même espace. Le critère de la méthode s'énonce alors ainsi :

$$FKM(\mathbf{A}, \mathbf{F}, \mathbf{U}|c, q) = \min_{\mathbf{U},\mathbf{A},\mathbf{F}} ||\mathbf{XA} - \mathbf{UF}||_F^2 \tag{4}$$

L'optimisation des critères (2) et (4) s'effectuent 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. L'algorithme s'arrête lorsque le critère ne décroît plus. Afin d'éviter la convergence vers un minimum local, les algorithmes nécessitent plusieurs initialisations.

Les deux approches ont des propriétés assez similaires, on pourra se référer à [12] pour une étude comparative. Notons que [2] propose une combinaison de ces deux approches, appelée *Generalized Reduced Clustering*, via la définition d'un critère d'optimisation exprimé comme combinaison linéaire des critères (3) et (4).

3 K-means sur données incomplètes

Comme évoqué en introduction, la gestion des données manquantes en classification peut s'effectuer efficacement par les méthodes d'imputation multiple [11, 13] et les méthodes dites *directes*, qui consistent à adapter les méthodes de classification de façon à ce qu'elles s'accommodent des données manquantes.

Dans le cas de la méthode des K-means, la gestion des données manquantes par imputation multiple se déroule en trois étapes distinctes [7]. Dans un premier temps, le jeu de données est imputé M fois selon un modèle d'imputation respectant la structure en groupes des observations [14]. La deuxième étape consiste à appliquer l'algorithme des K-means sur chacun des tableaux imputés, fournissant ainsi un ensemble de M partitions. Enfin, la troisième étape consiste en l'agrégation de ces dernières.

Concernant les méthodes "directes", [8] ont proposé de reformuler le critère (1) optimisé dans la méthodes K-means, de façon à ce qu'il puisse être évalué sur des données incomplètes. Celui-ci est défini comme suit :

$$f(\mathbf{U}, \mathbf{C}) = \min_{\mathbf{U},\mathbf{C}} \|P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{UC})\|_F^2 \tag{5}$$

avec $\Omega \subset \{1, \ldots, I\} \times \{1, \ldots, J\}$ un sous-ensemble des indices correspondant aux entrées observées. L'opérateur de projection des matrices $I \times J$ sur l'ensemble d'indices Ω est défini comme suit :

$$[P_\Omega(X)]_{ij} = \begin{cases} x_{ij} & \text{si } (i,j) \in \Omega \\ 0 & \text{si } (i,j) \in \Omega^c \end{cases}$$

Ainsi, le critère (5) revient à rechercher la matrice d'appartenance $\mathbf{U}$ et les centroïdes associés, de façon à ce que les profils observés (et donc partiels) des individus soient les plus proches de ceux des centres associés. Parce qu'il ne porte donc pas sur les éléments manquants du profil, il est parfaitement calculable sur un jeu de données incomplet. La solution du critère peut être approchée par un algorithme de Majorization-Minimization (MM). Son principe consiste à alterner entre le K-means et l'imputation des observations incomplètes par les coordonnées de leur centroïde associé.

4 Reduced *K-pod*

Afin de proposer une nouvelle approche de classification en grande dimension sur données incomplètes, nous proposons un nouveau critère d'optimisation reprenant les critères précédents, spécifiques aux méthodes de subspace clustering et de classification sur données incomplètes. Plus précisément, celui-ci est défini selon les notations précédentes de la façon suivante :

$$f(\mathbf{U}, \mathbf{A}, \mathbf{F}) = \min_{\mathbf{U},\mathbf{A}} \|P_\Omega(\mathbf{X}) - P_\Omega(\mathbf{U}\mathbf{F}\mathbf{A}^T)\|_F^2 \tag{6}$$

Ce critère peut être exprimé comme une généralisation du critère (2) au cas incomplet. Son optimisation peut alors être effectuée à l'aide d'un algorithme MM où la fonction majorante g au point courant $\tilde{U}, \tilde{F}, \tilde{A}$ se définit selon :

$$g(\mathbf{U}, \mathbf{F}, \mathbf{A}|\tilde{\mathbf{U}}, \tilde{\mathbf{F}}, \tilde{\mathbf{A}}) \quad = f(\mathbf{U}, \mathbf{A}, \mathbf{F}) + \left\| P_{\Omega^c}(\mathbf{U}\mathbf{F}\mathbf{A}^\mathbf{T}) - P_{\Omega^c}(\tilde{\mathbf{U}}\tilde{\mathbf{F}}\tilde{\mathbf{A}^\mathbf{T}}) \right\|_F^2$$

et se minimise selon un algorithme RKM, assurant ainsi la convergence monotone de l'algorithme Reduced k pod (1).

Algorithm 1 Reduced *K-pod*

Entrées : $\mathbf{X}$ un tableau de données incomplet, c un nombre de classes, q la dimension du sous-espace

Initialiser $\mathbf{F}^{(0)}$ et $\mathbf{A}^{(0)}$ par ACP itérative, puis mettre définir un tableau imputé $\mathbf{X}^{(0)}$ selon $\mathbf{X}^{(0)} \leftarrow P_\Omega(\mathbf{X}) + P_{\Omega^c}(\mathbf{F}^{(0)}\mathbf{A}^{(0)\top})$

Pour ℓ **de 1 à** $L-1$

- $(\mathbf{U}^{(\ell+1)}, \mathbf{F}^{(\ell+1)}, \mathbf{A}^{(\ell+1)}) \leftarrow$ Reduced k-means$(\mathbf{X}^{(\ell)})$
- $\mathbf{X}^{(\ell+1)} \leftarrow P_\Omega(\mathbf{X}) + P_{\Omega^c}(\mathbf{U}^{(\ell+1)}\mathbf{F}^{(\ell+1)}\mathbf{A}^{(\ell+1)\top})$

Sorties : $\mathbf{U}^{(L)}$, $\mathbf{F}^{(L)}$, $\mathbf{A}^{(L)}$

Cet algorithme consiste tout d'abord en une ACP itérative [15] afin d'obtenir les q composantes principales $\mathbf{F}^{(0)}$ et vecteurs propres associés $\mathbf{A}^{(0)}$. Ceux-ci sont alors utilisés pour initialiser les valeurs manquantes de $\mathbf{X}$. Cette initialisation permet de tenir compte de la structure en q dimension des données. Par la suite, l'algorithme alterne un RKM, afin de minimiser la fonction de coût g, et à mettre à jour les données manquantes de façon à pouvoir définir la valeur de la nouvelle fonction majorante au point courant. Ces étapes sont répétées jusqu'à atteindre un nombre prédéfini d'itérations L, ou arrêtées dès lors que le critère optimisé (6) ne décroît plus.

5 Simulations

5.1 Plan de simulation

Afin d'évaluer les performances de la méthode présentée, tant en termes d'identification de la partition que du sous-espace, nous nous plaçons dans le cadre d'un modèle de RKM (équation 3). 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 :

$$\mathbf{X} = \mathbf{UFA}^T + \mathbf{E}_{\text{sim}} \tag{7}$$

Pour cela, nous nous sommes inspirés du plan de simulation de [16]. 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, 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 $[-15, 15]^q$. $\mathbf{A}$ est ensuite construite de la sorte :

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}^{*T} & 0^T_{q\times(p2+p3)} \end{bmatrix}$$

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 $\mathcal{N}(0, \Sigma_J)$, avec Σ_J :

$$\Sigma_J = \begin{bmatrix} I_{p1} & 0_{p1\times p2} & 0_{p1\times p3} \\ 0_{p2\times p1} & \Sigma_{p2} & 0_{p2\times p3} \\ 0_{p3\times p1} & 0_{p3\times p2} & I_{p3} \end{bmatrix}$$

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 chacun des scénario, 200 jeux de données sont générés, en se basant sur des matrices $\mathbf{U}$, $\mathbf{F}$ et $\mathbf{A}$ fixes et en faisant varier $\mathbf{E}_{\text{sim}}$.
Pour chaque jeu de données, des données manquantes sont générées selon un mécanisme MCAR et MAR pour différents taux de données manquantes (5%, 15% et 25%). Sur les jeux de données complets, deux approches sont utilisées : RKM et l'approche tandem combinant l'ACP et le K-means sur les q premières composantes principales. Pour RKM, les paramètres utilisés sont $c = 8$, $q = 2$. Pour l'*approche tandem*, les mêmes paramètres sont appliqués. Ensuite, pour les jeux de données incomplets, trois approches sont envisagées : Reduced *K-pod*, l'équivalent de l'*approche tandem* adaptée aux données incomplètes et le *K-pod*. L'*approche tandem* adaptée aux données incomplètes commence par une ACP itérative suivie d'une étape de K-means. Les mêmes paramètres que pour les données complètes sont utilisés pour la *tandem approche* et le Reduced *K-pod* : $c = 8$, $q = 2$. Pour le *K-pod*, le paramètre c est fixé à 8.

Les performances des algorithmes sont mesurées selon l'indice de Rand ajusté (ARI) [4] entre les différentes partitions obtenues et la véritable partition des jeux de données. Pour Reduced K Means et Reduced *K-pod*, nous utiliserons le coefficient de congruence pour

comparer les matrices de *loadings* obtenues par les algorithmes et les matrices de *loadings* utilisées pour générer les données.

5.2 Résultats

Les Figures 1 et 2, représentent respectivement les ARI moyens sur les 200 jeux de données de ces deux scénarii ou $c = 8$, $n = 400$, $q = 2$ et $p_1 = p_2 = p_3 = 5$ ou 10.
Ici, nous constatons, dans un premier temps, une différence entre les ARI moyens entre les deux scénarii (avec un ARI moyen à 0.35 (Figure 1) et un ARI moyen 0.9 (Figure 2)) . Cela pourrait s'explique par une meilleure séparabilité des classes pour le deuxième scénario, observé sur le premier plan factoriel. Dans un second temps, on observe naturellement que plus le taux de données manquantes augmente, plus les performances des méthodes s'accommodant des données incomplètes diminuent. Pour la Figure 2, on constate que bien que les performances du Reduced *K-pod* décroissent plus rapidement, l'ARI du Reduced *K-pod* reste supérieure à celle de la tandem approche adaptée aux données manquantes. Dans les deux cas, le *K-pod* ne performe pas aussi bien que les deux autres méthodes. Cela peut s'expliquer par la structure en groupe qui se trouve dans un espace réduit ainsi que par l'initialisation des données manquantes par la moyenne des variables utilisée par la méthode du *K-pod*.

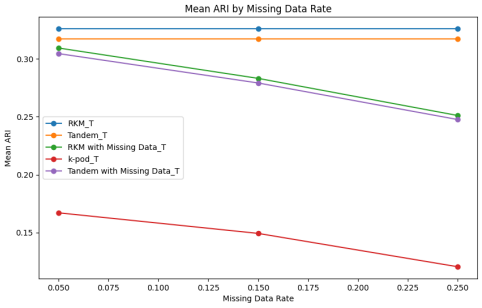

Figure 1: ARI moyen sur 200 jeux de données par taux de données manquantes (MCAR, $p_1 = p_2 = p_3 = 5$)

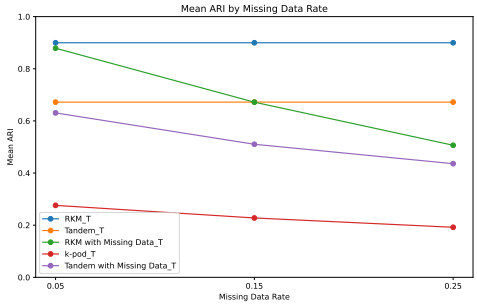

Figure 2: ARI moyen sur 200 jeux de données par taux de données manquantes (MCAR, $p_1 = p_2 = p_3 = 10$)

En ce qui concerne la Figure 3, on remarque que pour les différents taux de données manquantes, RKM performe en moyenne mieux que la *tandem approche* adaptée aux données manquantes pour chacun des jeux de données.

D'abord, on observe que la méthode Reduced *K-pod* a des valeurs d'ARI globalement supérieures à celles obtenues par une *approche tandem* avec ACP itérative.

En comparant les ARI obtenus selon Reduced *K-pod* et *K-pod*, tout comme précédemment, on observe que ces derniers sont plus élevés pour l'approche proposée. Cette différence plus marquée, peut s'expliquer par la prise compte de la structure en sous-espace. Cela peut

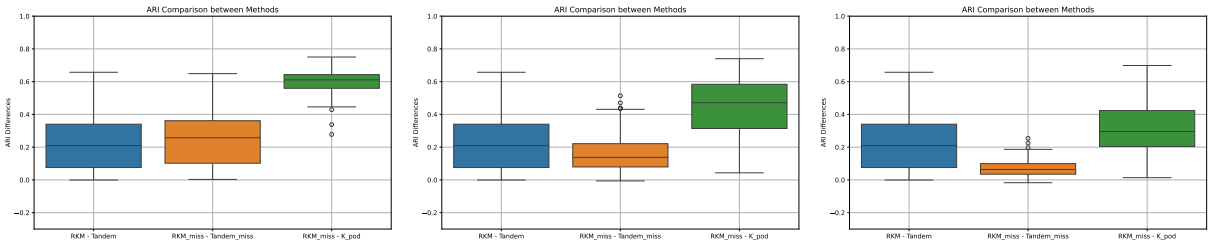

Figure 3: Distribution des différences d'ARI entre les différentes méthodes pour 5% (à gauche), 15% (au milieu) et 25% de données manquantes (à droite) selon un mécanisme MCAR, ($p_1 = p_2 = p_3 = 10$)

également s'expliquer par l'initialisation des données manquantes. En effet, la méthode du *K-pod* utilise une initialisation des valeurs manquantes par la moyenne des variables, ce qui peut altérer la structure en groupes des données.

6 Conclusion

Nous présentons ici une nouvelle méthode de classification sur données incomplètes dans le cadre d'un grand nombre de variables. Cette méthode s'appuie sur la formulation d'un critère tenant compte ces deux caractéristiques. L'algorithme d'optimisation associé converge de façon monotone. Cette première étude par simulation permet de mettre en évidence les meilleures performances de l'approche Reduced *K-pod* par rapport aux méthodes concurrentes (*K-pod* et *approche tandem*). Comme attendu, on constate une détérioration des performances avec le taux de données manquantes.

Cependant, ces performances sont à relativiser dans la mesure où le nombre de variables considéré dans cette simulation reste modeste. Une étude complémentaire est actuellement menée dans ce sens. D'autres travaux méthodologiques sont également en cours pour la prise en compte des données mixtes, à travers l'adaptation de la méthode des K-prototypes à la classification en grande dimension [10].

References

[1] Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. SIGKDD Explor., 6:90–105, 2004.

[2] Michio Yamamoto and Heungsun Hwang. A general formulation of cluster analysis with dimension reduction and subspace separation. Behaviormetrika, 41(1):115–129, 2014.

[3] Daniela M Witten and Robert Tibshirani. A framework for feature selection in clustering. Journal of the American Statistical Association, 105(490):713–726, 2010.

[4] Lawrence Hubert and Phipps Arabie. Comparing partitions. Journal of classification, 2:193–218, 1985.

[5] Geert De Soete and J Douglas Carroll. K-means clustering in a low-dimensional euclidean space. In New approaches in classification and data analysis, pages 212–219. Springer, 1994.

[6] Maurizio Vichi and Henk AL Kiers. Factorial k-means analysis for two-way data. Computational Statistics & Data Analysis, 37(1):49–64, 2001.

[7] Vincent Audigier and Ndèye Niang. Clustering with missing data: which equivalent for rubin's rules? Advances in Data Analysis and Classification, pages 1–35, 2022.

[8] Jocelyn T Chi, Eric C Chi, and Richard G Baraniuk. k-pod: A method for k-means clustering of missing data. The American Statistician, 70(1):91–99, 2016.

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

[10] Rabea Aschenbruck, Gero Szepannek, and Adalbert FX Wilhelm. Imputation strategies for clustering mixed-type data with missing values. Journal of Classification, 40(1):2–24, 2023.

[11] Donald B Rubin. Inference and missing data. Biometrika, 63(3):581–592, 1976.

[12] Marieke E Timmerman, Eva Ceulemans, Henk AL Kiers, and Maurizio Vichi. Factorial and reduced k-means reconsidered. Computational Statistics & Data Analysis, 54(7):1858–1871, 2010.

[13] R. Little and D. Rubin. Statistical Analysis with Missing Data. Wiley series in probability and statistics, New-York, 2002.

[14] Vincent Audigier, Ndèye Niang, and Matthieu Resche-Rigon. Clustering with missing data: which imputation model for which cluster analysis method? arXiv preprint arXiv:2106.04424, 2021.

[15] Julie Josse, François Husson, and Jérôme Pagès. Gestion des données manquantes en Analyse en Composantes Principales. Journal de la société française de statistique, 150(2):28–51, 2009.

[16] Yoshikazu Terada. Strong consistency of reduced k-means clustering. Scandinavian Journal of Statistics, 41(4):913–931, 2014.