Clustering on incomplete data using ***clusterMI***

V. Audigier

CNAM, CEDRIC-MSDMA, Paris

10èmes Rencontres R, June 13th, 2024

Clustering

Data $X_{n\times p} = (x_{ij})$ a mixed data set

Each individual i belongs to a unique cluster $\mathcal{C}_i \in \{1, ..., K\}$.

Aim identify $\mathcal{C}_i$ for each i using individual profiles $(x_i)_{1\leq i\leq n}$

Methods

Distance-based

- k-means
- k-prototypes partitioning
- hierarchical clustering
- partitioning around medoids (pam)

Model-based

- gaussian mixture models
- latent class models
- model-based clustering of mixed-type data
- mixture of multivariate t-distributions

Clustering with missing values

However, X is frequently **incomplete**... $x_i = (x_i^{obs}, x_i^{miss})$

Ad-hoc methods

removing incomplete observations/variables

Advanced methods

Direct methods

- k-means (Wagstaff, 2004; Honda et al., 2011; Chi et al., 2016)
- gaussian mixture (Miao et al., 2016; Marbac et al., 2019; McCaw et al., 2022)
- k-prototypes (Aschenbruck et al., 2022)

Multiple Imputation (MI)

- a three-stage approach
- references: Faucheux et al. (2020); Bruckers et al. (2017); Basagana et al. (2013); Audigier and Niang (2022)
- could be used for any clustering method

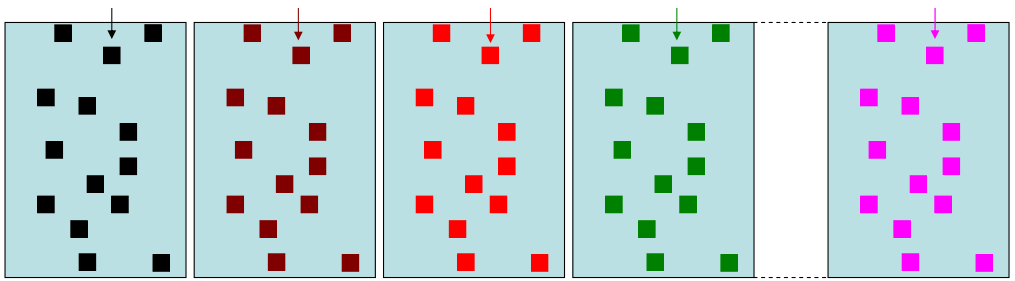
CRAN R packages

		ClustImpute, kpodclustr	MGMM, mixture, MixtureMissing	RMixtComp, MixAll, VarSelLCM	miclust	clusterMI
Clustering methods	kmeans	✓	✗	✗	✓	✓
	pam	✗	✗	✗	✗	✓
	hc	✗	✗	✗	✗	✓
	mixture	✗	✓	✓	✗	✓
Variable type	numeric	✓	✓	✓	✓	✓
	categorical	✗	✗	✓	✗	✓
	mixed	✗	✗	✓	✗	✓
Missing data	mcar	✓	✓	✓	✓	✓
	mar	✓	✓	✓	✓	✓
	mnar	✗	✗	✗	✗	✗
K	automatic ?	✗	✓	✓	✓	✓

Multiple imputation

1. Generate a set of M parameters $(\zeta_m)_{1 \leq m \leq M}$ of an imputation model to generate M plausible imputed data sets

$$P\left(X^{miss} | X^{obs}, \zeta_1\right) \quad \ldots \quad \ldots \quad \ldots \quad P\left(X^{miss} | X^{obs}, \zeta_M\right)$$

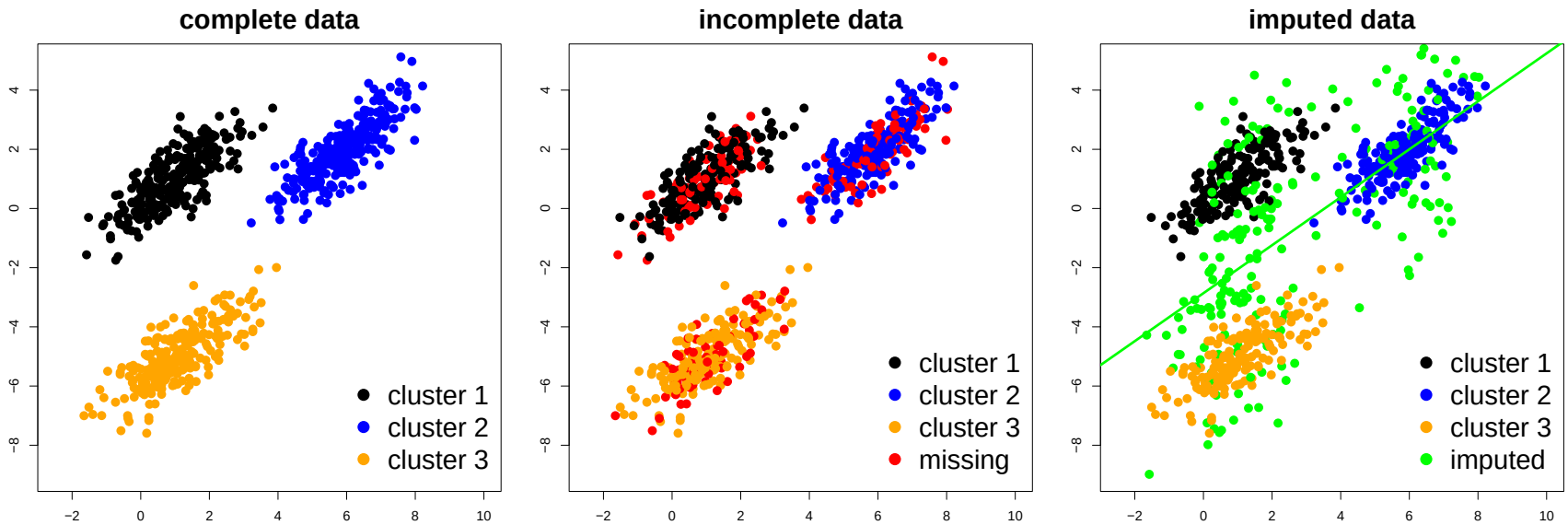


2. Apply the cluster analysis to each imputed data set: Ψ_m, U_m

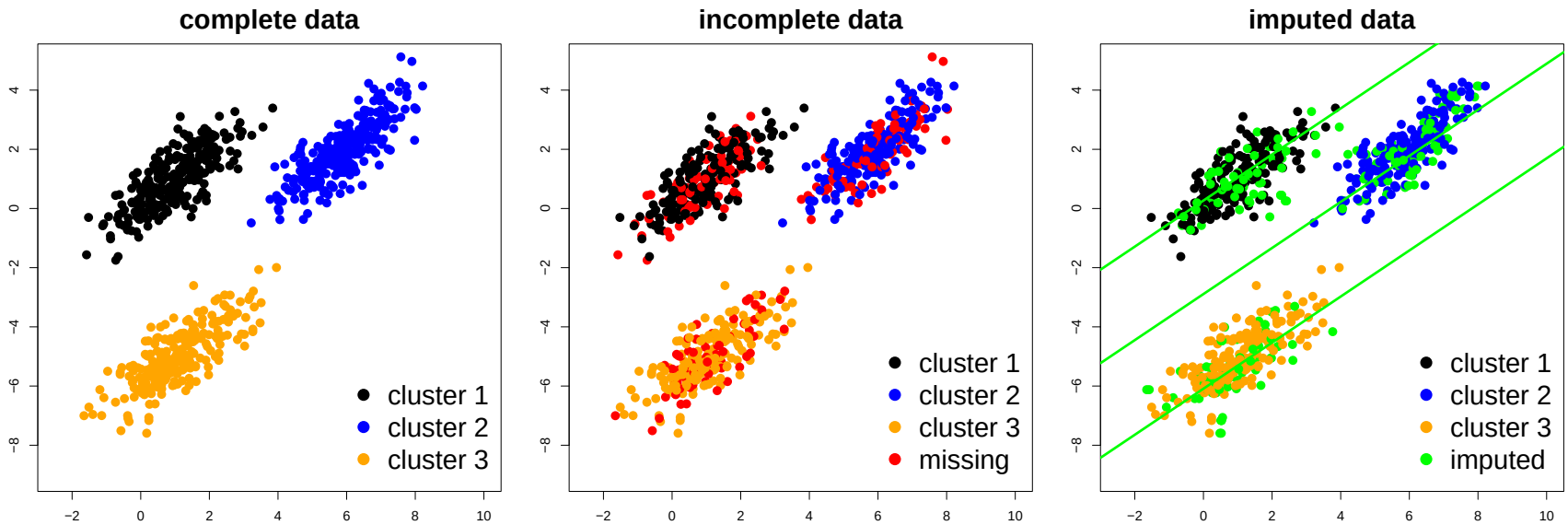
3. Combine the results to obtain one partition and one associated instability measure

Outline

Imputation model for clustering: the issue



Imputation model for clustering: the solution



Imputation methods in *clusterMI*

Joint Modelling

One joint distribution

- JM-DP (***DPimputeCont***, ***NPBayesImputeCat***)
- JM-GL (***mix***)

Fully Conditional Specification

A conditional distribution for each variable

- FCS-homo: linear model including a fixed effect for the cluster
- FCS-hetero: linear mixed effect model (***micemd***)

Properties

	JM-DP	JM-GL	FCS-homo	FCS-hetero
fast	✓	✓	✗	✗
heteroscedasticity	✓	✗	✗	✓
flexibility	✗	✗	✓	✓

```
1 imputedata(data.na, method = "JM-GL", nb.clust = 3)
```

Flexibility of FCS methods

Conditional imputation models can be easily modified

- to address specific data
  - non-gaussian distributions (Morris et al., 2014)
  - mixed data
  - complex relationships (Doove et al., 2014)

```
imputedata(data.na, method = "FCS-homo", nb.clust = 3,
           method.mice = c("pmm", "logreg", "rf", ...))
```

- to improve sparsity
  - penalized regression (Zahid and Heumann, 2019)
  - variable selection (Bar-Hen and Audigier, 2022)

```
imputedata(data.na, method = "FCS-homo", nb.clust = 3,
           method.mice = "lasso.norm",
           predictmat = matrix(...))
```

Variable selection (Bar-Hen and Audigier, 2022)

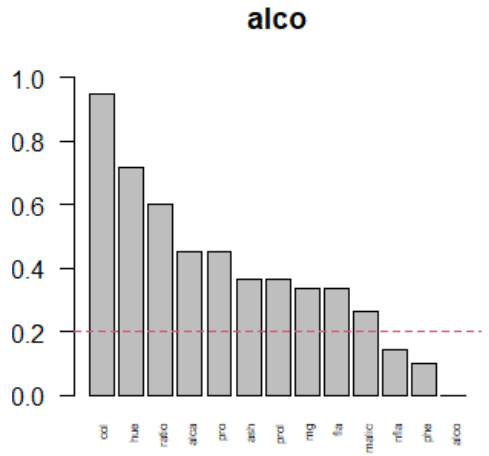
Tune a conditional imputation model for variable X_j

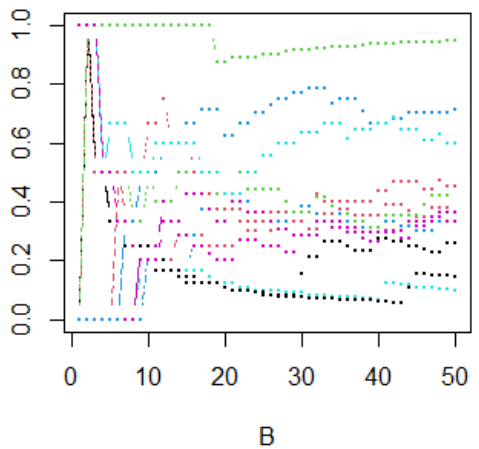
Repeat B times

- Sample k variables among $p-1$
- Handle missing values by JM-GL
- Apply a selection procedure on the subset

Compute

$$\frac{\#\text{ times the variable } X_j \text{ is selected}}{\#\text{ time } X_j \text{ is present in the instances}}$$

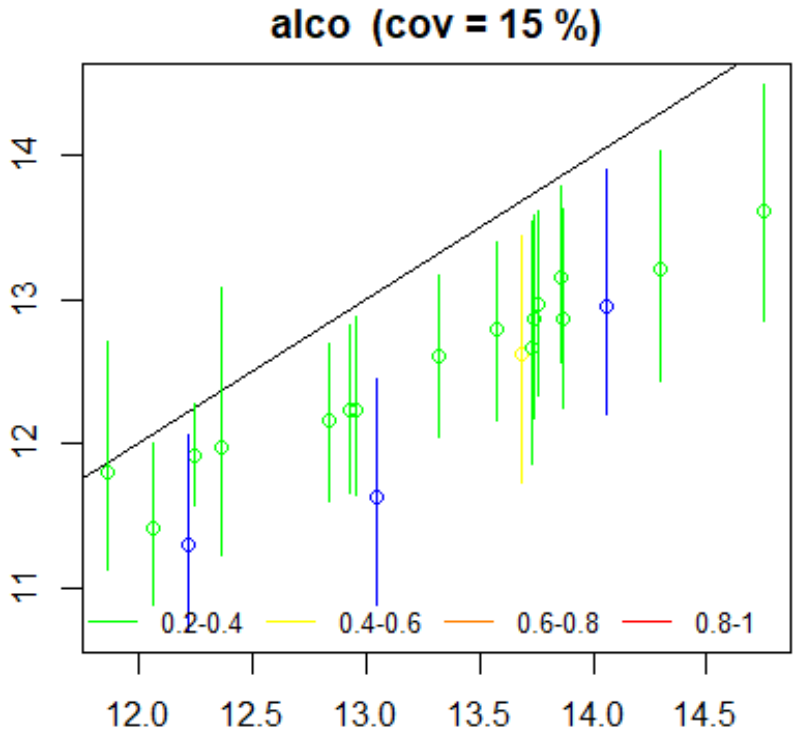





```
res.varsel <- varselbest(res.imputedata = res.imp.FCS,
                         listvar = "alco", nnodes = 20)
predictmat <-  res.varsel$predictormatrix
res.chooser <- chooser(res.varsel = res.varsel)
```

Overimputation (Blackwell et al., 2015)

1. perform multiple imputation
2. remove several observed values for a given variable
3. build 90% intervals using the percentile method




```
res.imp.over <- imputedata(data.na,
                           nb.clust = 3,
                           m = 200)
res.over <- overimpute(res.imp.over,
                       nnodes = 20,
                       plotvars = "alco",
                       plotinds = seq(nrow(data.na)))
```

Outline

Analysis

Apply the cluster analysis to each imputed data set

- various pre-implemented methods from ***fpc***
  kmeans, pam, clara, agnes or mixture, cmeans
- or any other method, e.g. reduced kmeans

Obtain $(\Psi_m)_{1\leq m\leq M}$ and compute $(U_m)_{1\leq m\leq M}$ (Fang and Wang, 2012)

- generate C bootstrap pairs $\left(X_c, \tilde{X}_c\right)_{1\leq c\leq C}$ from X
- perform cluster analysis from $\left(X_c, \tilde{X}_c\right)_{1\leq c\leq C}$ to obtain $\left(\Psi_c, \tilde{\Psi}_c\right)$
- classify individuals of X from Ψ_c and $\tilde{\Psi}_c$ to obtain $\left(\Psi'_c, \tilde{\Psi}'_c\right)$
- assess the instability $U = \frac{1}{C}\sum_{c=1}^{C} \frac{\delta(\Psi'_c, \tilde{\Psi}'_c)}{n^2}$

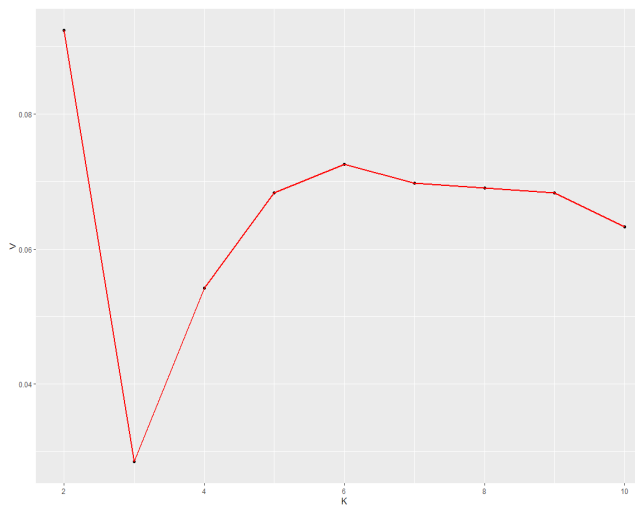


Figure: Instability (U) according to the number of clusters (K)

Pooling

Partitions pooling from $(\Psi_m)_{1\leq m\leq M}$ using NMF

Principle

- $(\Gamma_m)_{1\leq m\leq M}$ connectivity matrices associated to $(\Psi_m)_{1\leq m\leq M}$
- $\bar{\Gamma} = \frac{1}{M}\sum_{m=1}^{M}\Gamma_m$

$$\underset{\Gamma\in\mathcal{G}}{argmin} \parallel \Gamma - \bar{\Gamma} \parallel_F^2$$

Properties

- can be solved using various NMF algorithms
- monotone convergence
- no label switching problem, various choices for K_m are available

Instabilities pooling from $(U_m)_{1\leq m\leq M}$

$$T = \frac{1}{M}\sum_{m=1}^{M} U_m + \frac{1}{M^2}\sum_{m=1}^{M}\sum_{m'=1}^{M}\delta\left(\Psi_m, \Psi_{m'}\right)/n^2$$

Pooling in *clusterMI*

Using the R function clusterMI

```
> res.pool <- clusterMI(res.imp,
                        method.clustering = "kmeans",
                        instability = TRUE
                        nnodes = nnodes)
> names(res.pool)
[1] "part"        "instability" "call"
> names(res.pool$instability)
[1] "U"    "Ubar" "B"    "Tot"
```

```
choosenbclust(res.pool, grid = 2:5)
```

Or at hand for applying any clustering method, e.g. reduced kmeans

```
library(clustrd)
res.ana <- lapply(res.imp$res.imp,
                  FUN = cluspca,
                  nclus = 3, ndim = 2, method= "RKM")
res.ana <- lapply(res.ana, "[[", "cluster")
res.pool <- fastnmf(res.ana, nb.clust = 3)$cluster
```

Conclusion

A **R package** for clustering with missing values

- good performances
- flexible imputation methods
- a complete vignette

Specificities

- based on multiple imputation
- use dedicated imputation methods
- perform pooling using NMF
- compute instability measure to choose K

Perspectives

- parallel computing for FCS imputation
- add new clustering methods (dbscan, tclust, spectral clustering, ...)
- investigate imputation methods for MNAR mecachnisms

References I

Aschenbruck, R., Szepannek, G., and Wilhelm, A. F. X. (2022). Imputation strategies for clustering mixed-type data with missing values. *Journal of Classification*, 40(1):2–24.

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

Bar-Hen, A. and Audigier, V. (2022). An ensemble learning method for variable selection: application to high-dimensional data and missing values. *Journal of Statistical Computation and Simulation*, 0(0):1–23.

Basagana, X., Barrera-Gomez, J., Benet, M., Anto, J. M., and Garcia-Aymerich, J. (2013). A Framework for Multiple Imputation in Cluster Analysis. *American Journal of Epidemiology*, 177(7):718–725.

Blackwell, M., Honaker, J., and King, G. (2015). A unified approach to measurement error and missing data: Overview and applications. *Sociological Methods and Research*, pages 1–39.

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

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

References II

Doove, L., van Buuren, S., and Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, 72:92–104.

Fang, Y. and Wang, J. (2012). Selection of the number of clusters via the bootstrap method. *Comput. Stat. Data Anal.*, 56(3):468–477.

Faucheux, L., Resche-Rigon, M., Curis, E., Soumelis, V., and Chevret, S. (2020). Clustering with missing and left-censored data: A simulation study comparing multiple-imputation-based procedures. *Biometrical Journal*, n/a(n/a).

Honda, K., Nonoguchi, R., Notsu, A., and Ichihashi, H. (2011). Pca-guided k-means clustering with incomplete data. In *2011 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2011)*, pages 1710–1714.

Marbac, M., Sedki, M., and Patin, T. (2019). Variable selection for mixed data clustering: application in human population genomics. *Journal of Classification*, pages 1–19.

McCaw, Z. R., Aschard, H., and Julienne, H. (2022). Fitting gaussian mixture models on incomplete data. *BMC bioinformatics*, 23(1):1–20.

Miao, W., Ding, P., and Geng, Z. (2016). Identifiability of normal and normal mixture models with nonignorable missing data. *Journal of the American Statistical Association*, 111(516):1673–1683.

References III

Morris, T. P., White, I. R., and Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC medical research methodology*, 14(1):75.

Wagstaff, K. (2004). Clustering with missing values: No imputation required. In Banks, D., McMorris, F. R., Arabie, P., and Gaul, W., editors, *Classification, Clustering, and Data Mining Applications*, pages 649–658, Berlin, Heidelberg. Springer Berlin Heidelberg.

Zahid, F. M. and Heumann, C. (2019). Multiple imputation with sequential penalized regression. *Statistical Methods in Medical Research*, 28(5):1311–1327. PMID: 29451087.