IMPUTATION OF MIXED DATA WITH MULTI-LEVEL SVD

{ GENEVIÈVE ROBIN, FRANÇOIS HUSSON, JULIE JOSSE AND BALASUBRAMANIAN NARASIMHAN }
POLYTECHNIQUE, AGROCAMPUS, STANFORD

MOTIVATIONS

Analysis of the **Traumabase**® data. It includes 7495 major trauma patients distributed across 8 hospitals in the Paris area and 244 pre and post-hospital measurements. Major trauma is a public health challenge and a major source of mortality and handicap around the world. The project has several aims including

1. Monitor major trauma
2. Improve decision-making process & patients care
3. Provide real-time decision support.

Some scientific challenges are due to the nature of the data, which is **heterogeneous** (quantitative and qualitative variables), contains **missing values** and has a **multi-level** structure (patients within hospitals).

Centre	Lung X-ray	Mechanism	Lactate delay	Rea. time
Percy	NORM	AVP bike	NA	NA
Pitié S.	NORM	AVP ped.	NA	2
Beaujon	NA	Fall	0	56
Beaujon	NA	AVP car	NA	1
Henri M.	NA	AVP car	NA	1
Beaujon	ANORM	Fall	3	40

PROPOSAL

- **Handle missing values** with an imputation (completion) method dedicated to multi-level mixed data;
- Based on multi-level SVD method [4];
- Can be applied with small/large dimensions (n«p, n»p) and large numbers of categories for the factor variables;
- Computation can be distributed;
- Comparison to existing imputation methods for mixed data [1], [3];
- Application to the Traumabase data set.

MULTI-LEVEL IMPUTATION WITH PCA [4, 2] - QUANTITATIVE DATA

- Rationale: decompose variance into BETWEEN groups and WITHIN groups variances and perfom SVD on both parts.
- $Y \in \mathbb{R}^{n\times p}$: $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_N \end{pmatrix}$.
- N groups of size n_k for $1 \leq k \leq N$
- $Y = (y_{k,j,i_k})$, group k, var. j, ind. i_k.

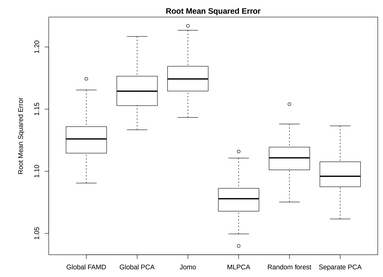
$$y_{k,j,i_k} = y_{.,j,.} + (y_{k,j,.} - y_{.,j,.}) + (y_{k,j,i_k} - y_{k,j,.}),$$
$$Y = \bar{Y} + Y_b + Y_w$$
$$= 1_n m' + U_b V_b' + U_w V_w' + E.$$

- Missing values: $\Omega \in \{0,1\}^{n\times p}$ observations mask.
- Least squares problem (under linear constraints)
$$\underset{m,U_b,U_w,V_b,V_w}{\operatorname{argmin}} \left\|\Omega \odot (Y - 1_n m' - U_b V_b' - U_w V_w')\right\|^2,$$
$\Rightarrow$ **Iterative imputation algorithm** [2].

Input: Y, Ω, q_b, q_w
Output: m, U_b, U_w, V_b, V_w
Initialize: $\hat{Y} = Y \odot \Omega + 1_n m_0' \odot (1 - \Omega)$

1. Estimate $\hat{Y}_b$: PCA of the matrix of group means $\sum_{k=1}^{N} 1_{n_k} \frac{1}{n_k}(1_{n_k}' \hat{Y} - m_0')$
2. Estimate $\hat{Y}_w$: PCA of the centered matrix $\hat{Y} - 1_n m_0' - \hat{Y}_b$
3. Impute $\hat{Y} = Y \odot \Omega + \hat{Y} \odot (1 - \Omega)$; $m = n^{-1} 1_n' \hat{Y}$

Repeat steps 1, 2, 3 until convergence



MULTI-LEVEL IMPUTATION WITH MCA - QUALITATIVE DATA

- $Y \in \mathbb{R}^{n\times p}$ categorical data set.

$$Y = \begin{pmatrix} 1 & 1 \\ 2 & 3 \\ 1 & 2 \end{pmatrix} \iff Z = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \end{pmatrix},$$

- π vector of proportions;
- Z_b Between part: for each k

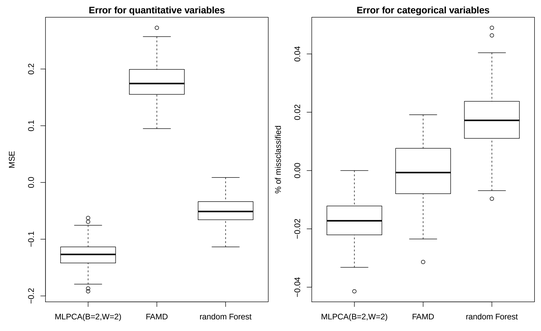
$$Z_{b,k} = (1/n_k) 1_{n_k} (1_{n_k}' Z_k - \pi);$$

- Z_w Within part $Z_w = Z - 1_n \pi' - Z_b$,
$$\Rightarrow Z = 1_n \pi' + Z_b + Z_w.$$

Iterative imputation with **MCA**.

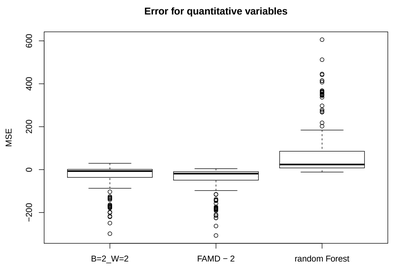
1. Estimate $\hat{Z}_b$ with MCA of Z_b
2. Estimate $\hat{Z}_w$ with MCA of Z_w
3. Impute $\hat{Z} = Z \odot \Omega + \hat{Z} \odot (1 - \Omega)$; $\pi = n^{-1} 1_n' \hat{Z}$

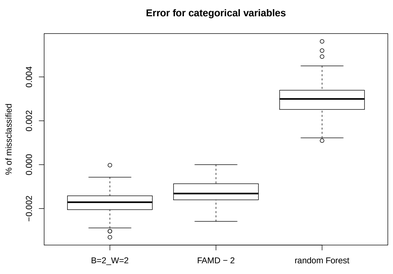
Using multi-level MCA and PCA combined we can impute mixed-data.



TRAUMABASE - FIRST RESULTS

- Select 8 variables based on doctors recommandations and exploratory data analyses
- Hospital, Lung X-ray, Pelvis X-ray, Mechanism, CGR 24h, Delay normalization lactate, Last socio-pro activity, Time of departure for scanner or op., Time in reanimation.
- Introduce 5% of MCAR missing values.
- Impute with three methods: multi-level SVD, FAMD [1] (component method for mixed data), Missforest [3].
- 200 replications.





- Missforest has highest errrors in this example.
- Multi-level SVD and FAMD perform similarly on quanti. variables.
- Multi-level SVD improves slightly the imputation of quali. variables.

REFERENCES

[1] F. Husson and J. Josse. Handling missing values in multiple factor analysis. *Food Quality and Preference*, 30(2):77 – 85, 2013.

[2] J. Josse, M. E. Timmerman, and H. A. Kiers. Missing values in multi-level simultaneous component analysis. *Chemometrics and Intelligent Laboratory Systems*, 129:21 – 32, 2013. Multiway and Multiset Methods.

[3] D. Stekhoven and P. Bühlmann. Missforest - nonparametric missing value imputation for mixed-type data. *Bioinformatics*, 28:113–118, 2012.

[4] M. E. Timmerman. Multilevel component analysis. *British Journal of Mathematical and Statistical Psychology*, 59(2):301–320, 2006.

CONCLUSION & FUTURE RESEARCH

- Two SVD (between & within) to impute multi-level mixed data;
- Improves on competitors (computational time & size of data);
- Can preserve confidentiality.
- Distribution of the method and deployment on the distcomp platform [**?**];
- Influence of group effective;
- Confidence in the predicted values: multiple imputation.