Apprentissage Machine / Statistique

# Imputation de Données Manquantes

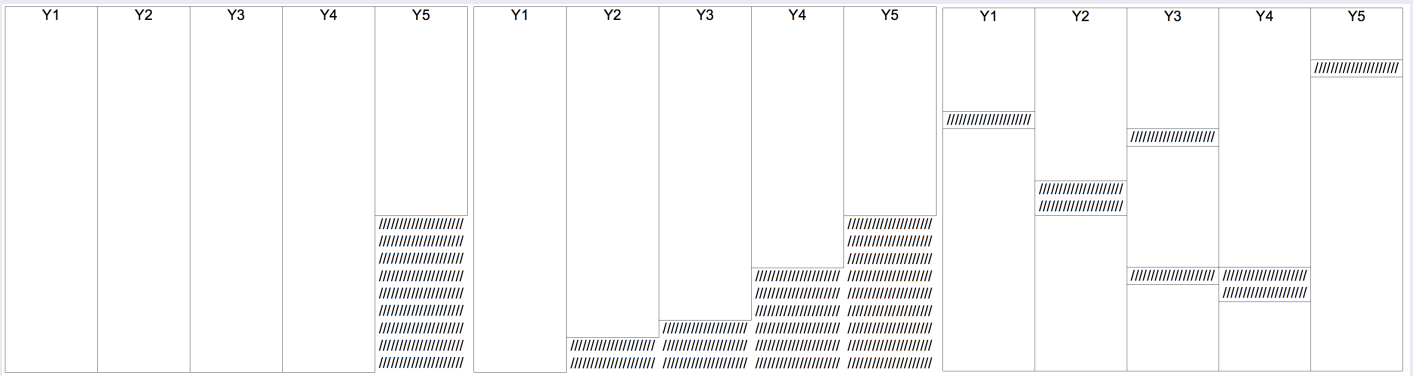
PHILIPPE BESSE

INSA – Dpt GMM
Institut de Mathématiques – ESP

## Types de données manquantes

- Missing completly at random (MCAR)
- Missing at random (MAR) : absence liée à une ou plusieurs autres variables
- Missing not at random (MNAR) : absence dépend de la variable
- Type des variables : quantitatives ou qualitatives

## Répartition univariée, monotone, arbitraire



## Analyse sans imputation

- Suppression
  - Lignes : perte de précision, biais si non MCAR
  - Colonnes
- Méthodes tolérantes
  - CART : divisions de substitution
  - NIPALS en PLS (imputation implicite)
  - XGBoost

## Imputations unidimensionnelles

- Complétion stationnaire : plus fréquente, dernière valeur, *Last Observation Carry Forward (LOCF)*
- Combinaison linéaire : moyenne, médiane

## Imputations par régression

- $k$ plus proches voisins
- Régression, régression locale (LOESS)
- NIPALS (PLS, PCA) ou SVD

## Imputation par SVD

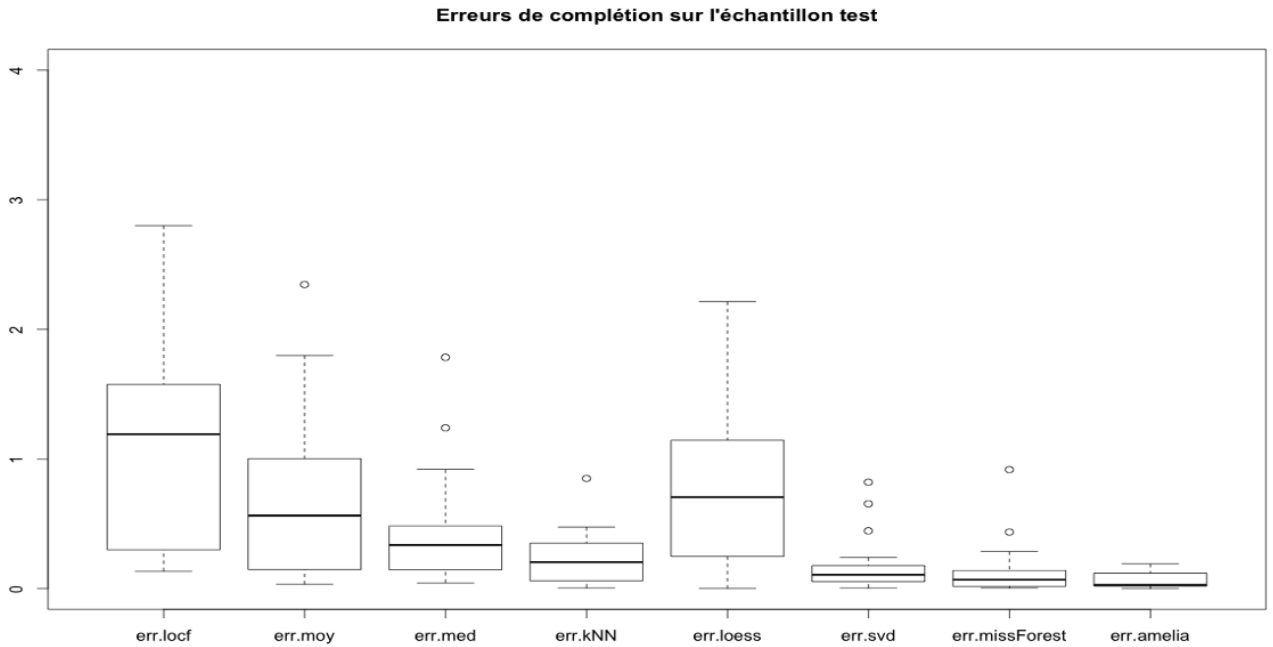
- Matrice $\mathbf{Y}^0$ complétée de $\mathbf{Y}$ par des moyennes
- **Tant que** $\| Y^i - Y^{i+1} \| / \| Y^i \| < \epsilon$
  - $\widehat{Y}$ : SVD tronquée (rang $J$) de $\mathbf{Y}^i$ centrée
  - $\mathbf{Y}^{i+1}$ est $\mathbf{Y}$ complétée par les éléments de $\widehat{Y}$

## Imputation multiple : Amelia II

- Imputer plusieurs valeurs
- Combinaison des résultats
- Réduction du bruit, mesure d'incertitude
- Hypothèse de normalité
- Algorithme EMB (Expectation Minimisation et Bootstrap)

## Imputation par forêts aléatoires

1. Complétion "naïve" des valeurs manquantes.
2. $k$ indices des colonnes de $Y$ triées par quantité croissante de valeurs manquantes ;
3. **Tant que** $\gamma$ n'est pas atteint **faire**
   1. $Y_{imp}^{old}$ = matrice précédemment imputée
   2. **Pour** $s$ dans $k$ **faire**
      1. Ajuster $y_{obs}^{(s)} \sim x_{obs}^{(s)}$ par forêt aléatoire
      2. Prédire $y_{mis}^{(s)}$ avec les régresseurs $x_{mis}^{(s)}$
      3. $Y_{imp}^{new}$ : nouvelle matrice complétée par les $y_{mis}^{(s)}$
   3. mettre à jour le critère $\gamma$




*EBP - Erreurs de complétion sur un échantillon test de* $10\%$