Imputation Multiple avec des Variables-Flags

Modou DIA, CEPS-INSTEAD,
Differdange (LUXEMBOURG)

JMS 14-16 mars 2005

Quel type de non-réponse étudié ?

- Non-réponse partielle :
  - Un item manquant pour une unité répondante ou une grappe répondante
  - Elle concerne une variable continue : le salaire brut

Mécanismes de non-réponse envisagés

- Missing Completely At Random (MCAR)
- Missing At Random (MAR)

MCAR

Soient:

- X la variable explicative
- Y la variable dépendante
- Définition: La non-réponse est de type MCAR si la probabilité de réponse est indépendante de X et de Y

MAR

Définition:

La Probabilité de réponse est dépendante de X, mais pas de Y

Pourquoi une méthode d'imputation multiple

- Les méthodes d'imputation dite simple ne donnent qu'une solution parmi plusieurs envisageables
- Elles sont incapables de restituer la variabilité inhérente à l'incertitude de la distribution des valeurs imputées
- Grâce à l'imputation multiple, il est possible de calculer la variance liée à l'imputation

En quoi consiste l' Imputation Multiple

- Les observations manquantes sont remplacées par $m > 1$ échantillons de données
- Analyse de chaque pseudo-échantillon
- Combinaison des m estimateurs

AVANTAGE:

la variabilité de l'échantillonnage est prise en compte par la variance liée à l'estimation

Les principaux modèles d'imputation multiple

- Les méthodes basées sur l'Estimateur de Maximum de Vraisemblance classique (EMV)

- Les méthodes de type Data Augmentation (DA) basées sur une inférence bayésienne

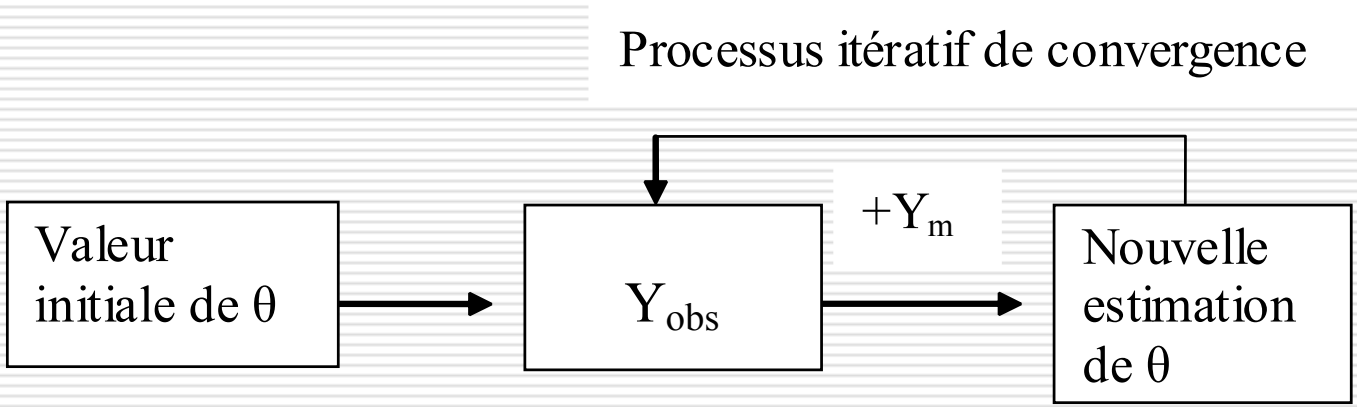
Les principaux algorithmes utilisés

- L'algorithme Expectation-Maximization (EM) pour les modèles de type EMV classique

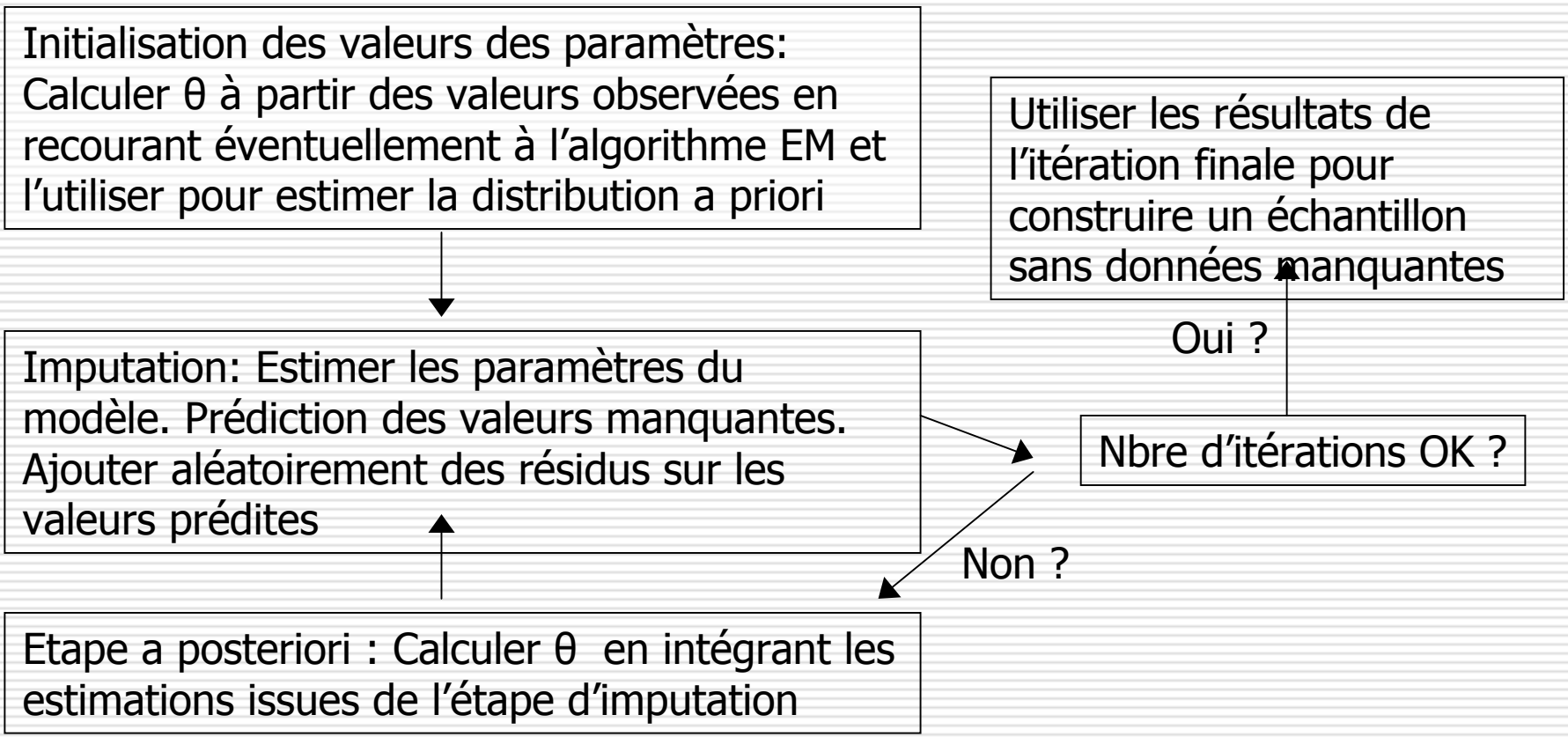
- L'algorithme Markov Chain Monte Carlo (MCMC) pour les modèles de type DA

Algorithme EM

Soient :
Y la variable à imputer ;
Y_{obs} les observations renseignées ;
Y_m les observations manquantes ;
θ les paramètres du modèle



Algorithme du MCMC



Logiciel ou programme utilisé

- IVEware –(Imputation Variance Estimation) développé par RAGHUNATHAN T.E. , LEPKOWSKI J.M., VAN HOEWYK J., and SOLENBERGER P. (1999) de L'Université de Michigan
  http://www.isr.umich.edu/src/smp/ive/

- C'est un ensemble de macros fonctionnant sous SAS

Caractéristiques d'IVEware

- C'est une régression généralisée séquentielle
- Soit X, l'ensemble de départ des variables explicatives sans données manquantes
- Soient Y_i i= 1, 2,.., k, les k variables dépendantes classées selon l'ordre croissant de leur taux de données manquantes
- Cela donne $[Y_1/X]$ $[Y_1/X,Y_2]$ …. $[Y_k/X,Y_2..Y_{k-1}]$

Les 4 principales familles de modèles d'IVEware

- une régression linéaire si Y_i est continue
- une régression logistique si Y_i est binaire;
- une régression polytomique si Y_i est catégorielle ;
- un modèle suivant la loi de Poisson si Y_i est une variable discrète finie;

Hypothèses sur les flags (1)

- Soient une variable Y = (y_1, y_2, … y_i…. Y_n) et F_i, le flag associé à chaque observation de y_i telles que :
- -F_i=1, si y_i est observé ;
- -F_i=2, si y_i est manquant ;
- -F_i=3, si y_i est « non-concerné » par exemple dans le cas d'un salaire pour un chômeur (se) ou un(e) retraité(e)

Hypothèses sur les flags (2)

- Soient :
- r, le nombre d'observations à valeurs renseignées ;
- m, le nombre d'observations à valeurs manquantes ;
- nc, le nombre d'observations non-concernées ;
- tels que : n = r+m+nc, nombre total d'observations

Calcul de la moyenne quand les observations “non-concernées” sont nulles

pour le calcul initial de θ, on a :

$$\overline{Y} = 1/(r + nc) * \sum_{Fi=1}^{n} y_i$$

et pour les valeurs suivantes de θ :

$$\overline{Y} = (1/(r + m + nc)) * (\sum_{Fi=1}^{n} y_i + \sum_{Fi=2}^{n} y_i)$$

Calcul de la moyenne quand les observations "non-concernées" sont non nulles

les valeurs de $\bar{Y}$ pour le calcul des valeurs initiales ou non-initiales de θ seront respectivement :

$$\overline{Y} = 1/r \sum_{F_i=1}^{n} y_i$$

et

$$\overline{Y} = (1/(r+m+nc))*(\sum_{F_i=1}^{n} y_i + \sum_{F_i=2}^{n} y_i + \sum_{F_i=3}^{n} y_i)$$

Conséquences sur les valeurs du paramètre θ

- Une sous-estimation de la moyenne
- Une sous-estimation de la variance

C'est le résultat inverse s'il s'agit d'une variable revenu formulé négativement comme une perte pour les indépendants

Source et but de l'application

- Source : Enquête Panel Socio-Economique Liewen zu Lëtzebuerg (PSELL3), première vague EU-SILC

- Variable à imputer : salaire brut

Variables du modèle de regression pour variable continue

- Variables continues : le temps de travail, le ratio salaire net-salaire brut, la pension de retraite
- Variables catégorielles : le secteur public ou privé, le niveau d'éducation, la CSP, le sexe, la classe d'imposition, la carte d'impôt
- Variables discrètes : l'âge, l'expérience professionnelle, le nombre d'enfants

Options et Contraintes du modèle

- 0.55 ≤ Ratio salaire net-brut ≤1
- P5 ≤ Salaire brut ≤ P95
- P5 ≤ Pension de retraite ≤ P95
- Nombre d'imputations =50
- MINRSQD =0.01 (Minimum marginal R-Squared)

Moyenne et écart-type des valeurs collectées durant l'enquête et des valeurs imputées

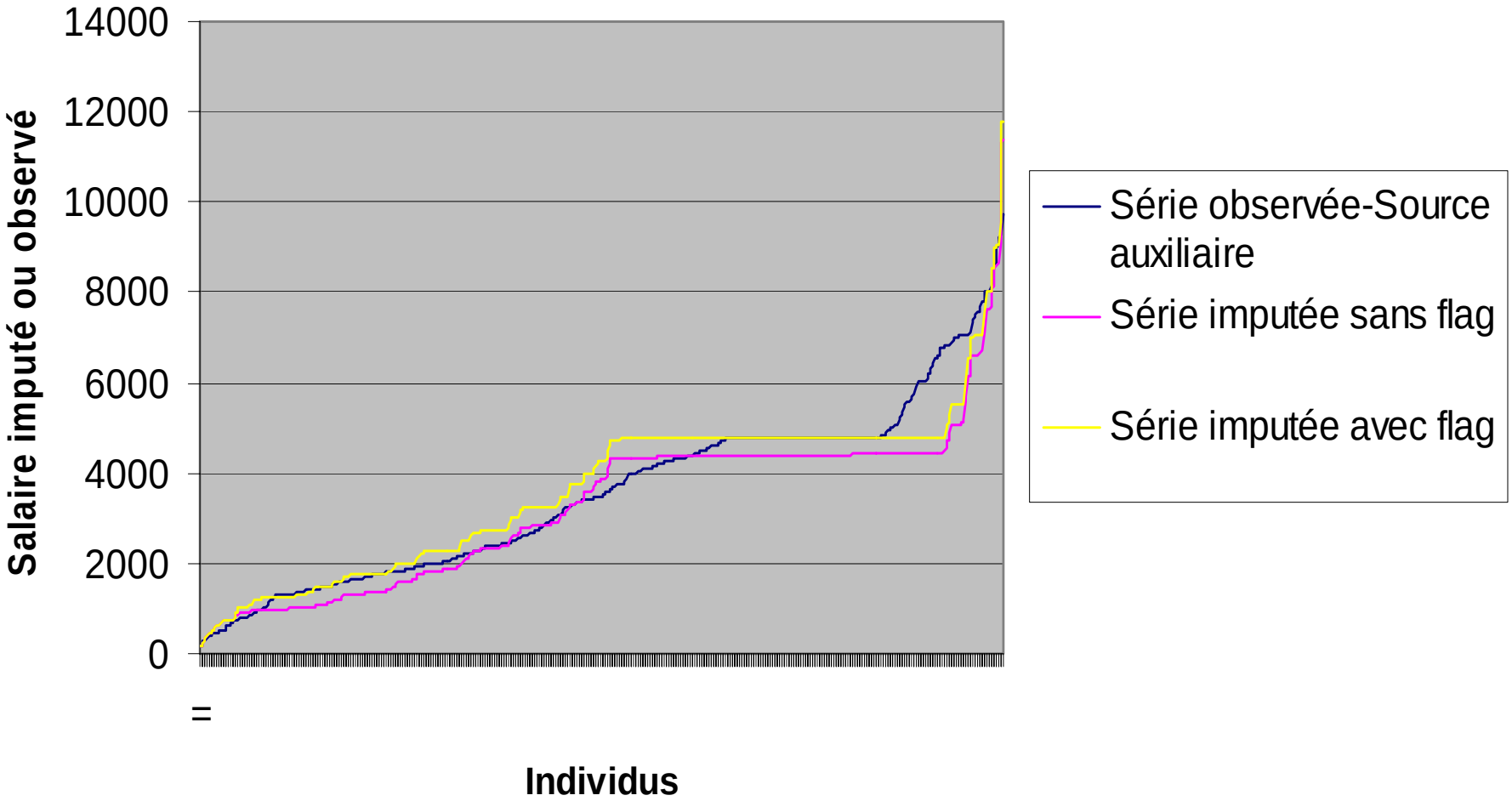
Les moyennes des valeurs imputées sans flags sont sous-estimées car les individus les plus qualifiés et les plus expérimentés sont sur-représentés pami les observations manquantes

FLAG \ type de données		Observées	Imputées
Sans flag	Moyenne	3259.69	3201.02
Avec flag	Moyenne	3259.69	3557.32
Sans flag	Ecartype	2114.16	1765.41
Avec flag	Ecartype	2114.16	1821.09

Comparaison des données imputées (avec ou sans flag) avec celles de la source auxiliaire disponibles

- Confirmation de la tendance observée

Comparaison des données imputées (avec ou sans flag) avec celles de la source auxiliaire



Erreur entre les valeurs imputées et les valeurs observées dans la source externe

- L'estimation est plus précise pour les valeurs imputées avec flag

Flag	Nb_imputations	Erreur_imputation	Nb_observations
NON	50	2087.45	656
OUI	50	1871.20	656

Conclusion

- Si le choix d'un modèle d'imputation revêt un caractère capital
- Les conditions de sa mise en œuvre n'en restent pas moins importantes
- L'utilisation du programme IVEware, avec flag ou sans flag, en est une illustration .

- Merci de votre attention
- Des questions ?