IMPUTATION MULTIPLE DE DONNÉES CATÉGORIELLES : UNE APPROCHE BASÉE SUR UN MODÈLE MULTINORMAL LATENT.

Anne De Moliner (*), Philippe Périé (**)

(*) ENSAE
(**) TNS Sofres, Direction Scientifique

Introduction

Dans les grandes enquêtes, les unités échantillonnées ne répondent pas toutes complètement au questionnaire. Certaines n'y répondent pas du tout et d'autres ne répondent qu'à certaines questions. La qualité de la réponse et le taux de remplissage dépendent des conditions de passation du questionnaire et en particulier de la longueur de celui-ci.

Une approche pour traiter ce genre de non réponse est *l'imputation multiple des données manquantes.* C'est une méthode statistique pour l'analyse des données incomplètes originellement proposée par Rubin (1987). Plusieurs déclinaisons de la méthode existent qui diffèrent par les modèles spécifiés, leurs hypothèses et la façon de générer les imputations.

Dans ce papier, nous proposons une nouvelle approche pour le traitement des données catégorielles (en particulier binaires) de grande dimensionnalité, portant sur des évènements rares. Ce type de données est plus fréquent qu'on peut le penser à première vue : des choix de consommations entre plusieurs références de produits sur une période, l'audience au sens de la lecture dernière période sur les titres de la presse, un questionnement d'attribution sur une batterie de plusieurs dizaines d'items, etc. … Il pose en plus pas mal de problèmes au praticien du fait des effectifs et donc des croisements très faibles.

Notre proposition reprend le cadre normal multivarié décrit par Schafer (1997) et s'appuie sur la spécification d'un modèle latent, pour lequel les valeurs observées vont déterminer des troncatures. Les valeurs imputées sont donc tirées conditionnellement à ces troncatures.

Nous utilisons la méthode GHK (Geweke, Hajivassiliou, Keane) pour tirer dans des lois multinormales tronquées, et l'approche de Berens (2008) pour la détermination du modèle gaussien latent, ensuite c'est l'approche de Schafer qui est utilisée. Les jeux de données d'exemple sont tirés de données réelles et simulées d'évènements rares de grande dimension (200 variables, 40000 observations). Sur ces données nous avons simulé plusieurs taux et arrangements de valeurs manquantes.

Avant de décrire notre méthode, nous donnons d'abord quelques repères sur le cadre d'analyse et les concepts les plus importants. Cette partie nous permettra de mieux situer l'algorithme que nous proposons dans la palette des outils disponibles, puis nous étudierons en exemple un test pour une méthode de questionnement en blocs incomplets.

1. Cadre de référence

1.1. Notations

Soit Y un vecteur de variables d'intérêt incomplètes de dimension *k* : $Y = (Y_i, ..., Y_k)$. Soit Y_j une des *k* variables incomplètes $(j = 1, ..., k)$. Les parts observées et manquantes des Y_j sont notées Y_j^{obs} et Y_j^{manq} respectivement. Ainsi $Y^{obs} = (Y_1^{obs}, ..., Y_k^{obs})$ et $Y^{manq} = (Y_1^{manq}, ..., Y_k^{manq})$ indiquent les parts

observée et manquante des données Y. Soit $Y_{-j} = \left(Y_i,...,Y_{j-1},Y_{j+1},...Y_k\right)$ l'ensemble des variables dans *Y* hors Y_j . Soit R le vecteur des indicatrices de réponse sur Y : $R = \left(R_1,...,R_k\right)$ où R_j est le vecteur des indicatrices de réponse pour Y_j avec $R_j = 1$ si Y_j est observé et $R_j = 0$ sinon. On définit aussi $R_{-j} = \left(R_1,...,R_{j-1},R_{j+1},...R_k\right)$. Soit $X = \left(X_1,...,X_l\right)$ un ensemble de *l* covariables complètement observées sur les mêmes sujets.

La théorie de l'imputation multiple pour traiter les données manquantes exige que l'imputation soit faite conditionnellement au plan de sondage. Une façon de le faire est d'inclure les caractéristiques du plan de sondage dans l'ensemble des covariables. Pour le moment, afin éviter des complexités de notation, nous supposerons que les observations sur Y, X et R correspondent à un tirage aléatoire simple dans la population d'intérêt.

1.2. Principe de l'imputation multiple

Le principe de l'imputation multiple est de remplacer une table contenant des valeurs manquantes par m tables complètes, avec m≥2, obtenues en tirant les valeurs manquantes dans une distribution spécifiée préalablement. On obtient les résultats recherchés en combinant les résultats obtenus séparément sur chacune des tables complétées. On peut ensuite calculer la variance des estimateurs ainsi produits grâce à des formules préétablies, on intègre ainsi l'incertitude due aux valeurs manquantes.

On distingue depuis Rubin (1987) [15] trois principales phases dans une analyse :

La première est la plus complexe : elle consiste en la modélisation et l'estimation, puis la génération de données plausibles, appelées imputations, pour les valeurs manquantes. Le rôle de l'étape de modélisation est de spécifier une distribution jointe pour les données, ensuite on cherche à obtenir la distribution bayésienne prédictive *a posteriori* des valeurs manquantes conditionnellement aux données observées. L'estimation consiste à calculer la distribution des paramètres de cette distribution, de manière à ce que l'on puisse tirer aléatoirement dedans. Cette étape permet de remplacer la table contenant les valeurs manquantes par m tables complètes dans lesquelles les valeurs manquantes sont remplacées par des tirages aléatoires dans une distribution de valeurs plausibles. Le nombre d'imputations m varie typiquement entre 3 et 10.

La deuxième étape consiste à analyser chaque fichier de données imputées avec des techniques statistiques classiques qui vont estimer les variables d'intérêt. On obtient donc m analyses (au lieu d'une) qui vont différer seulement parce que les imputations diffèrent.

La troisième étape combine les estimations en une seule, en décomposant les variations dans et entre les imputations. Sous certaines conditions assez souples, cette étape permet d'obtenir des estimations valides qui incorporent l'incertitude due aux données manquantes dans les intervalles de confiance.

1.3. Modèles et hypothèses

1.3.1. Hypothèse d'ignorabilité des valeurs manquantes

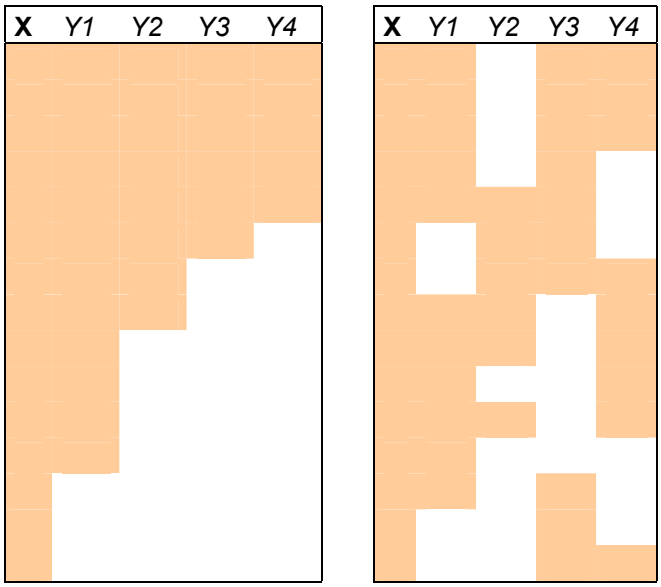
L'information sur *Y* qui est présente dans *X* et *R* est résumée par la distribution conditionnelle P(Y/X, R). Les observations avec des données manquantes, c'est-à-dire avec R=0, ne fournissent pas d'information sur P(Y/X, R), ce n'est possible qu'avec les valeurs observées, donc l'analyse ne se fait qu'avec P(Y/X, R=1). Mais le problème est bien de tirer des imputations à partir de la distribution P(Y/X, R=0) qu'il faut spécifier. La procédure conventionnelle est de poser que P(Y|X, R = 0) = P(Y|X, R = 1), ce qui correspond à l'hypothèse que le mécanisme de non-réponse est ignorable (Rubin, 1987, pp. 51–53).

1.3.2. Dispositions des valeurs manquantes

Si les valeurs manquantes sont disposées de manière monotone, alors des imputations multivariées peuvent être obtenues par une séquence de tirages dans des lois univariées. Supposons que les variables $Y = (Y_1,..., Y_k)$ soient ordonnées de manière monotone de telle sorte que pour $(j = 1,...., k-1)$ toutes les observations avec des valeurs manquantes pour Y_j aient aussi des valeurs manquantes pour $Y_{>j}$. Si les paramètres $\theta_1,..., \theta_k$ des modèles d'imputation sont *a priori* indépendants, alors on peut tirer dans la loi multivariée en utilisant la séquence suivante de tirages univariés :

$$P\left(Y_1^{manq} / X, \theta_1\right)$$
$$P\left(Y_2^{manq} / X, Y_1^*, \theta_2\right)$$
$$P\left(Y_3^{manq} / X, Y_1^*, Y_2^*, \theta_2\right)$$
...
$$P\left(Y_k^{manq} / X, Y_1^*,..., Y_{k-1}^*, \theta_k\right)$$

Où Y_j^* est la j[ème] variable après imputation. On répète la séquence *m* fois pour obtenir les réplications, ce n'est pas la peine d'itérer. Dans les cas où il est impossible, même après réarrangements, d'obtenir un schéma monotone, il faut une méthode vraiment multivariée.



arrangement monotone ***arrangement général***

1.3.3. Modèles pour imputations multivariées

Lorsqu'on est en arrangement non monotone, on doit avoir recours aux méthodes d'imputation multivariées. La complexité des lois manipulées interdit alors les estimations par calcul analytique, et l'on doit avoir recours à des méthodes itératives. On peut distinguer en imputation multivariée deux approches : une basée sur spécification de la loi jointe (JM, Joint Modelling), et une basée sur la spécification de distributions conditionnelles (FCS, Fully Conditionnal Spécification), que l'on appelle aussi séquences de régressions.

1.3.3.1. Méthodes basées sur la spécification d'une loi jointe

Ces méthodes partent en spécifiant une densité paramétrique $P(Y, X, R/\theta)$, pour les données Y, X et R sachant θ. En spécifiant une distribution *a priori* pour θ, on définit des sous-modèles spécifiques

pour chaque arrangement de valeurs manquantes, dans lesquels les imputations sont tirées sous l'hypothèse d'ignorabilité. Selon le type de données manipulées, on dispose de plusieurs modèles. **On utilise pour cela les méthodes MCMC** : dans une simulation MCMC, on construit une chaîne de Markov qui est assez longue pour que la distribution des éléments se stabilise en une distribution stationnaire, qui est la distribution d'intérêt. En répétant assez longtemps les étapes de la chaîne, on peut simuler des tirages dans la distribution d'intérêt. Dans l'inférence bayésienne, l'information sur les paramètres inconnus est exprimée sous la forme de la distribution *a posteriori*. On la calcule avec la formule de Bayes : $p(\theta/y) = \frac{p(y/\theta)p(\theta)}{\int p(y/\theta)p(\theta)d\theta}$

En appliquant les méthodes MCMC, on peut facilement simuler complètement la distribution *a posteriori* et obtenir des estimateurs des paramètres d'intérêt.

Toutefois, dans beaucoup de problèmes de données incomplètes, la probabilité *a posteriori* des données observées $P(\theta/Y_{obs})$ est très difficile à simuler. Pour cela, on *augmente* les données observées Y_{obs} par une estimation (simulation) des données manquantes Y_{manq}, et la probabilité *a posteriori* des données complètes $P(\theta/Y_{obs,}, Y_{manq})$ est plus facile à simuler.

Le principe d'augmentation des données peut être appliqué en inférence bayésienne en répétant les étapes suivantes :

1. La phase d'imputation (I-step) : étant donné une estimation des paramètres $\hat{\theta}$, cette phase simule (tire) les valeurs manquantes pour chaque observation indépendamment. C'est-à-dire que pour chaque individu i on tire ses valeurs de la distribution conditionnelle de $Y_{manq(i)} / Y_{obs(i)}$
2. La phase de tirage des paramètres (Posterior step ou P-step) : étant donné un échantillon complet, cette phase simule la distribution *a posteriori* des paramètres $\hat{\theta}$. Ces nouvelles estimations seront utilisées à nouveau dans la phase I-step suivante. Sans information sur les paramètres, on utilise une distribution *a priori* dite non informative, mais on peut introduire dans cette phase de l'information *a priori*.

Les deux étapes sont itérées assez longtemps pour que les estimations se stabilisent. Ainsi avec un paramètre dans l'itération courante $\theta^{(r)}$, le I-step tire $Y_{manq}^{(r+1)}$ dans $p(Y_{manq}/Y_{obs}, \theta^{(r)})$ et le P-step tire $\theta^{(r+1)}$ depuis $p(\theta/Y_{obs}, Y_{manq}^{(r)})$. Ceci crée une chaîne de Markov qui converge en distribution vers $p(Y_{manq}, \theta/Y_{obs})$. Une fois que l'on a convergé vers la distribution stationnaire, on simule des tirages approximativement indépendants des valeurs manquantes dans cette distribution. On valide la méthode en répétant ce processus en partant de valeurs initiales différentes.

Pour les données continues, la méthode d'imputation multiple basée sur la spécification d'un modèle normal multivarié (Schafer 1997) [17] est très largement utilisée. Ses avantages sont nombreux : la méthode est très largement disponible dans les logiciels commerciaux ou libres, elle est simple et efficace à implémenter car travailler avec la loi normale permet beaucoup de simplifications avec des routines d'algèbre linéaire très facilement accessibles.

Pour les données catégorielles ou mixtes, Schafer propose des modèles de type loglinéaire ou des modèles de type localisation générale multivariée, mais alors les algorithmes sont vraiment plus complexes et les rares solutions logicielles ne sont pas du tout utilisables pour des problèmes en taille réelle. Cette difficulté d'application et la relative robustesse du modèle normal a conduit à tenter de l'appliquer au traitement des données catégorielles. Mais même si de nombreuses études tendent à montrer que la méthode est relativement robuste aux écarts à la multinormalité, elle achoppe complètement pour des données catégorielles d'autant plus qu'elles portent sur des évènements rares (Horton, Lispsitz, Parzen 2003) [8].

1.3.3.2. Méthodes basées sur des tirages dans les lois conditionnelles

Une autre voie consiste à faire l'économie de la spécification explicite d'une loi jointe complexe, et de travailler avec une série de lois conditionnelles simples, une pour chaque variable traitée. En itérant les tirages dans les lois conditionnelles, on cherche à obtenir la loi jointe. Cette technique est connue sous le nom de séquences de régressions (Raghunathan et al. 2001) [12], ou Fully Conditional Specification (FCS, Stef van Bureen 2005) [20]. Cette approche offre une grande flexibilité pour des données d'enquête dans lesquelles on peut avoir des dépendances complexes (censures, croisements vides, …). Elle est facile à appliquer avec les solutions logicielles déjà développées (mais pas à reprogrammer). Toutefois ses propriétés statistiques sont difficiles à établir car l'existence de la loi jointe n'est pas garantie dans tous les cas. Les simulations faites par les auteurs tendent à montrer que l'approche se comporte très bien même dans ces cas, mais nous avons noté une sous estimation des corrélations dans les cas de données catégorielles de grande dimensionnalité sur des évènements rares.

1.3.3.3. Relations entre les deux méthodes

Dans certains cas simples, les deux approches coïncident. Par exemple si $L(\ )$ est une loi normale multivariée, alors toutes ses densités conditionnelles sont des combinaisons linéaires avec une erreur normale : en imputant avec une série de régressions linéaires, on réalise exactement la même chose qu'une imputation avec le modèle normal de Schafer. Un autre cas est celui d'un ensemble de variables binaires avec un modèle loglinéaire avec des interactions d'ordre 2. Dans ce cas les distributions conditionnelles $P(Y_i / Y_1,...,Y_{i-1},Y_{i+1},...,Y_k)$ sont des régressions logistiques. En imputant avec une série de régressions logistiques, on réalise la même chose qu'avec le modèle loglinéaire de Schafer dans lequel on aura fixé les interactions d'ordre supérieur à deux à zéro.

Ce dernier résultat est très important pour nous car il a déterminé la méthode à laquelle nous avons comparé notre méthode dans les simulations. Nous avons comparé notre méthode dans le cas de variables binaires au modèle loglinéaire de Schafer, en utilisant une solution logicielle plus simple et plus stable, basée sur l'approche FCS.

2. Notre modèle : MVN sur variables latentes

2.1. Les variables latentes

2.1.1. Spécifier des variables latentes

Nous faisons l'hypothèse qu'à nos variables binaires Y_i correspondent des variables latentes Y_i^* normales qui les déterminent. Mathématiquement, on postule donc l'existence d'un vecteur latent Y_i^* tel que : $Y_i = 0$ si $\mathrm{Y_i^*} < S_i$ et $Y_i = 1$ sinon avec $\mathrm{Y^*} \sim \mathrm{N}(\gamma, \Lambda)$. Où S est le vecteur des seuils, que l'on peut fixer à 0 sans perte de généralité pour des raisons d'identifiabilité.

2.1.2. Pourquoi travailler sur des variables latentes ?

D'abord, une spécification qui fait sens pour nos données et qui est assez classique en théorie des choix : on pose l'existence d'une fonction latente d'utilité, dont on n'observe que la concrétisation en termes de choix. Par exemple dans le cas où les variables binaires représentent le fait de consommer ou non un produit, la variable latente peut être l'utilité reçue de la consommation de ce produit : on ne le consomme que si l'utilité qu'on en retire est supérieure aux coûts.

Ensuite, les limites rencontrées sur les solutions logicielles testées : Nos données sont constituées de variables binaires de grande dimensionnalité, avec de très faibles occurrences, ce qui pose des problèmes numériques à beaucoup de logiciels. La première solution a été de tenter de travailler directement sur les variables binaires avec l'approche de Schafer, et de spécifier un modèle

loglinéaire. Même en fixant à zéro les interactions d'ordre supérieur à 2, on doit estimer pour k variables $k \times (k-1)$ coefficients sur des cellules presque vides, ce qui a été très au-delà des possibilités du logiciel CAT de J. Schafer sur S+. Ensuite en exploitant l'équivalence entre séquences de régressions logistiques et modélisations jointe loglinéaire, nous avons travaillé avec le logiciel IVEWARE de Raghunathan depuis SAS 9.2 et l'implémentation FCS dans SPSS 17. Les programmes ont bien tourné mais nous avons remarqué une sous-estimation systématique des interactions entre les variables.

La spécification de ces variables latentes normales nous permettait donc d'appliquer le modèle de Schafer dans le cadre normal, donc dans les conditions les plus simples à programmer, avec la bibliothèque de routines la plus large, ce qui laisse la place à l'optimisation numérique dans ces cas extrêmes.

2.1.3. Trouver les moments des variables latentes à partir des moments des variables binaires

Il nous faut maintenant trouver la loi de ces variables latentes en nous basant sur la partie observée directement : les variables binaires. Pour cela nous nous appuyons sur la méthode décrite dans l'article de Make, Berens et Ecker [6] expliquant comment déduire les moments de variables latentes normales multivariées à partir des moments des variables binaires correspondantes, sous réserve que ces variables latentes existent. En conservant les notations précédentes, on peut, sans perte de généralité, poser des variances unitaires i.e. $\Lambda_{i,i} = 1$. Soient *r* le vecteur des moyennes des variables binaires et Σ leur matrice de variance-covariance. On démontre aisément que les moments des variables latentes et observées sont liés par les relations suivantes :

$$r_i = \phi(\gamma_i)$$
$$\Sigma_{i,i} = \phi(\gamma_i)\phi(-\gamma_i)$$
$$\Sigma_{i,j} = \Psi(\gamma_i, \gamma_j, \Lambda_{i,j})$$

Avec $i \neq j$, $\Psi(x, y, \lambda) = \Phi(x, y, \lambda) - \Phi(x)\Phi(y)$. Φ étant la fonction de répartition d'une loi normale $N(0,1)$ et $\Phi(x, y, \lambda)$ la probabilité pour que deux variables normales de corrélation (covariance) λ soient supérieures à leurs seuils respectifs γ_i et γ_j. On peut donc déduire γ directement grâce à la relation $\gamma_i = \Phi^{-1}(r_i)$.

Pour trouver Λ, il faut résoudre le système d'équations $\begin{cases} \gamma_i = \Phi^{-1}(r_i) \\ \Sigma_{i,j} - \Psi(\gamma_i, \gamma_j, \Lambda_{i,j}) = 0 \end{cases}$ pour chaque couple (i, j), et cela nécessite que les solutions existent et soient uniques. On démontre que c'est le cas grâce au théorème des valeurs intermédiaires et à quelques calculs de probabilités, cf. [6] pour plus de détails sur les calculs.

2.1.4. Corriger les problèmes de matrices non définies positives

Malheureusement, il est possible dans certains cas que cette matrice Λ, obtenue coefficient par coefficient, ne soit pas une matrice de variance covariance, c'est-à-dire qu'elle ne soit pas définie positive du fait des arrondis cumulés. Pour corriger pratiquement ce problème il faut trouver une matrice définie positive la plus proche possible de la matrice obtenue.

Plusieurs méthodes existent, la plus utilisée est la méthode dite 'alternating projections' de Higham [5] Pour des raisons de rapidité d'exécution, nous avons utilisé une alternative plus immédiate dite 'square root' : soit $\hat{\Sigma}$ une matrice de covariance estimée. Si $\hat{\Sigma}$ n'est pas définie positive, pour obtenir une matrice définie positive, on peut considérer sa décomposition en racine carrée $\hat{\Sigma} = A^2$ avec $A = A_1 + iA_2$ (où A_1 et A_2 sont des matrices définies positives). On obtient un nouvel estimateur des

covariances avec $\hat{\Sigma}^* = A_1^2$ et $\hat{\Sigma}^*$ est définie positive. L'approximation obtenue est très proche de la matrice de départ, cette étape n'induit pas de biais notables dans nos résultats.

2.1.5. Tirage des variables latentes conditionnellement aux valeurs binaires

2.1.5.1. La loi normale tronquée multivariée

Maintenant que nous avons estimé les paramètres de la loi conditionnelle, il faut tirer des valeurs dans ces lois en tenant compte de l'information apportée par les valeurs binaires observées : on doit tirer dans (Y^*/Y) avec Y^* le vecteur latent. Ces variables suivent une loi normale tronquée multivariée, de densité : $f_Y(x) = C \times (2\pi)^{-N/2} |\Sigma|^{-N/2} \exp\{(x-\mu)\Sigma^{-1}(x-\mu)'\} I_R(x)$

où $I_R(x)$ est une indicatrice valant 1 si les contraintes imposées par les valeurs binaires sont respectées et 0 sinon : $R = \Pi_{i=1,..,N}[a_i, b_i]$

et $[a_i, b_i] = [0, \infty[\; si Y_i = 1 \text{ et } [a_i, b_i] =]-\infty, 0] \text{ sinon}$

et *C* est une constante de normalisation : $C = \int_{a_1}^{b_1}\int_{a_2}^{b_2}...\int_{a_3}^{b_3} f_{Y^*}(y)dy$

2.1.5.2. L'algorithme de tirage

Nous avons utilisé le simulateur GHK (du nom de Geweke, Hajivassiliou et Keane), qui permet de tirer dans une loi normale tronquée multivariée selon le principe suivant.
On veut tirer le vecteur *Y* tel que :

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \\ \\ Y_p \end{pmatrix} \sim TN(\gamma, \Lambda, A, B) \equiv N(\gamma, \Lambda) \, s.c. \, A < Y < B \quad (1)$$

Posons $U = (u_1, u_2, ...u_p)$ un vecteur suivant une loi normale multivariée centrée réduite et $\Lambda^{1/2}$ la décomposition de Choleski (triangulaire inférieure) de Λ (i.e. $\Lambda = \Lambda^{1/2} ' \Lambda^{1/2}$) dont les éléments sont les suivants : $\begin{pmatrix} l_{1,1} & 0 & 0 & 0 \\ l_{2,1} & l_{2,2} & 0 & 0 \\ ... & ... & l_{i,i} & 0 \\ l_{p,1} & l_{p,2} & ... & l_{p,p} \end{pmatrix}$

L'équation (1) équivaut à : $Y = \gamma + \Lambda^{1/2} U \sim N(\gamma, \Lambda)$ s.c.

$$\begin{pmatrix} \frac{a_1-\mu_1}{l_{1,1}} \\ \frac{a_2-\mu_2-l_{2,1}u_1}{l_{2,2}} \\ \dots \\ \frac{a_p-\mu_p-\Sigma_{i-1}^{p-1}l_{p,i}u_i}{l_{p,p}} \end{pmatrix} < \begin{pmatrix} u_1 \\ u_2 \\ \dots \\ u_p \end{pmatrix} < \begin{pmatrix} \frac{b_1-\mu_1}{l_{1,1}} \\ \frac{b_2-\mu_2-l_{2,1}u_1}{l_{2,2}} \\ \dots \\ \frac{b_p-\mu_p-\Sigma_{i-1}^{p-1}l_{p,i}u_i}{l_{p,p}} \end{pmatrix} \quad (2)$$

Comme on a arrangé les éléments de la matrice de corrélations en triangulaire inférieure il suffit pour tirer dans la loi multivariée de tirer les valeurs récursivement dans des lois normales tronquées univariées.

On commence donc par tirer u_1 selon $\mathrm{TN}\left(0,1,\frac{a_1-\mu_1}{l_{1,1}},\frac{b_1-\mu_1}{l_{1,1}}\right)$,

Puis u_2 selon $\mathrm{TN}\left(0,1,\frac{a_2-\mu_2-l_{2,1}u_1}{l_{2,2}},\frac{b_2-\mu_2-l_{2,1}u_1}{l_{2,2}}\right)$ et ainsi de suite….

Il ne nous reste alors plus qu'à calculer le vecteur *Y* grâce à la transformation $Y=\gamma+\Lambda^{1/2}U$.

2.1.5.3. Intégration des covariables : analyses factorielles

Les covariables X qui sont entièrement observées sont intégrées dans le modèle par le biais de leurs facteurs issus d'analyses factorielles. Les facteurs sont des combinaisons de variables orthogonales entre elles et centrées, donc faciles à manipuler, et qu'il est raisonnable de considérer comme normalement distribuées. Pour des variables continues, on réalise des analyses en composantes principales, pour des variables qualitatives, des analyses de correspondances multiples. On peut les réaliser en sous blocs de variables de même nature. Dans ce qui suit nous allons noter X, comme les données originales, les facteurs issus des analyses factorielles, pour indiquer que nous les prenons comme une transformation particulière des covariables de départ, et qu'ils jouent donc le rôle de covariables.

Cette intégration change légèrement la phase de tirage dans la loi normale tronquée multivariée, car il faut considérer les valeurs observées qui vont conditionner pour chaque individu les valeurs latentes tirées. On a donc une nouvelle matrice Λ de variance-covariance en ajoutant les facteurs X issus du bloc des covariables, et sa décomposition de Choleski est de la forme suivante :

$$\begin{pmatrix} \lambda_1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \dots & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \lambda_j & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \lambda_l & 0 & 0 & 0 & 0 \\ c_{1,1} & c_{1,2} & \dots & \dots & c_{1,l} & l_{1,1} & 0 & 0 & 0 \\ c_{2,1} & c_{2,2} & \dots & \dots & c_{2,l} & l_{2,1} & l_{2,2} & 0 & 0 \\ c_{j,1} & c_{j,2} & \dots & c_{j,j} & c_{j,l} & \dots & \dots & l_{j,j} & 0 \\ c_{p,l} & c_{p,2} & \dots & \dots & c_{p,l} & l_{p,1} & l_{p,2} & \dots & l_{p,p} \end{pmatrix}$$

La séquence de tirages dans les lois normales tronquées se fait comme précédemment, mais conditionnellement aux valeurs observées : les coordonnées des individus sur les facteurs, les variances et les covariances entre les facteurs et les vecteurs latents.

2.2. Algorithme MCMC et génération des imputations

On cherche donc à tirer les valeurs manquantes latentes dans leur loi jointe conditionnellement aux valeurs latentes associées aux valeurs observées, sachant que les paramètres de ladite loi sont inconnus. Pour cela, on va utiliser un échantillonneur de Gibbs. Le principe général du Gibbs Sampling est le suivant : on tire dans une loi jointe complexe par tirages successifs dans des lois conditionnelles plus simples. Ici nous allons donc tirer successivement les valeurs manquantes latentes conditionnellement aux paramètres et aux valeurs latentes des observées (C'est la phase d'imputation ou I-Step) puis les paramètres conditionnellement à l'ensemble des valeurs latentes (C'est la Posterior Step ou P-Step). La chaîne de Markov finit par converger.

2.2.1. Initialisation de l'algorithme

Il faut initialiser l'algorithme en générant un tirage initial des valeurs manquantes. Pour cela, on peut utiliser n'importe quelle technique d'imputation simple, comme le hot deck par exemple. Nous avons choisi de commencer par estimer les paramètres puis les moments de la loi normale multivariée sur la partie observée des données. On a ensuite tiré les valeurs latentes dans la loi normale multivariée ainsi définie, sans troncature puisque ces valeurs ne correspondent pas à des valeurs observées.

2.2.2. Répétition successive des I-Step et P-Step suivantes jusqu'à la convergence

2.2.2.1. Phase d'imputation : I-Step :

Il s'agit d'une version légèrement modifiée de la I Step de l'échantillonneur de Gibbs classique, notre I Step se compose de deux étapes

a) A chaque itération, on tire les valeurs latentes correspondant aux valeurs binaires observées conditionnellement à ces valeurs binaires, mais aussi aux paramètres et aux valeurs manquantes imputées. La loi de ces valeurs latentes est une loi normale tronquée conditionnelle :

$$\left(\mathrm{Y}^{*}_{\mathrm{obs}} \middle/ \mathrm{Y}^{*}_{\mathrm{manq}}, \mathrm{Y}_{\mathrm{obs}}, \mathrm{X}, \mu, \Sigma\right) \sim \mathrm{TN}\left(\mu_{\mathrm{cond}}, \Sigma_{\mathrm{cond}}, \mathrm{S}\right)$$

On rappelle que le vecteur des seuils *S* est constant égal à au vecteur nul.

Et $\mu_{cond} = \mu_{obs} + \Sigma_{obs,manq}{}^{-1}\Sigma_{manq}\left(Y_{manq} - \mu_{manq}\right)$ et $\Sigma_{cond} = \Sigma_{obs} - \Sigma_{obs,manq}\Sigma_{manq}{}^{-1}\Sigma_{manq,obs}$

Avec Σ_{obs} la partie de la matrice de variance covariance correspondant aux observées (X et Y_{obs}), Σ_{manq} la partie correspondant aux manquantes, et $\Sigma_{obs,manq}$ la partie représentant les corrélations entre observées (X et Y_{obs}) et latentes.

C'est cette phase qui a été rajoutée à l'échantillonneur de Gibbs classique. En effet, nous avons vu que les paramètres des lois normales tronquées sont estimés à partir des paramètres de la loi binaire, et comme l'estimation de ces paramètres change à chaque itération les valeurs de ces variables latentes observées sont stochastiques et dépendent du reste des estimations, il fallait donc les tirer à nouveau selon les nouveaux paramètres à chaque itération pour les inclure dans la chaîne de Markov afin d'obtenir sa convergence globale.

b) On tire ensuite les valeurs latentes des valeurs manquantes conditionnellement aux paramètres et aux valeurs latentes des variables observées.

$$\left(Y^{*}_{manq} \middle/ Y_{obs}, X, \mu, \Sigma\right) \sim N\left(\mu_{cond}, \Sigma_{cond}\right)$$

Avec $\mu_{cond} = \mu_{manq} + \Sigma_{obs,manq}{}^{-1}\Sigma_{manq}\left(y_{obs} - \mu_{obs}\right)$ *et* $\Sigma_{cond} = \Sigma_{manq} - \Sigma_{manq,obs}\Sigma_{mobs}{}^{-1}\Sigma_{obs,manq}$

2.2.2.2. P Step :

La P-Step est la phase de tirage des paramètres conditionnellement aux données.

On commence par tirer Σ dans une loi de Wishart inverse[1] :

$$\Sigma \sim W^{-1}(N+m,(N-1)S+L)$$

Avec *N* le nombre d'individus et $(N-1)S$ la matrice de corrélation empirique calculée à chaque itération. m et L sont des paramètres de la loi a priori, sur laquelle nous donnerons des explications plus bas.

On tire ensuite *μ* dans une loi normale : $(\mu/\Sigma, X, Y_{obs}, Y_{manq}) \sim N(\overline{Y}, N^{-1}\Sigma)$

Avec N le nombre d'individus, $\overline{Y}$ le vecteur des moyennes. Ceci correspond à l'utilisation d'un *a priori* non informatif sur les moyennes.

2.2.3. Utilisation d'un a priori

Au départ, nous avions postulé un *a priori* non informatif (de Jeffreys) pour les variances, mais notre algorithme ne débouchait pas sur la convergence de la distribution jointe mais au contraire conduisait à une sous-estimation des corrélations, augmentant avec le nombre d'itérations. Pour résoudre ce problème, nous avons spécifié une loi *a priori* de la matrice des corrélations, ce qui a permis de stabiliser l'inférence. On a donc choisi de passer d'un *a priori* non informatif à un *a priori* informatif. Plus précisément on postule la loi a priori de Σ, $\Sigma \sim W^{-1}(m, L)$

On pourra par exemple prendre pour paramètre L la corrélation obtenue sur une enquête précédente comportant les mêmes variables si elle est disponible. Le paramètre m quant à lui représente le poids de l'*a priori* face aux observations. Cette loi a priori peut être vue comme une manière d'incorporer une information *a priori* dans notre algorithme.

On pourrait éventuellement ajouter aussi un *a priori* sur les moyennes.

2.2.4. Convergence et récupération des imputations

On réitère le processus ci-dessus un nombre suffisant de fois. Les *i* premières valeurs ne sont pas prises en compte (c'est la *phase d'initialisation* ou *burn in)* car on considère que la convergence n'a pas encore eu lieu. A l'issue de cette période, on stocke un tirage toutes les *P* itérations, avec *P* suffisamment grand (usuellement entre 100 et 1000) ; ce *P* peut être choisi par exemple grâce à la statistique *R* de Gelman et Rubin. Les tirages peuvent être considérés comme indépendants s'ils sont suffisamment espacés.

Pour avoir *m* imputations, on doit donc effectuer *i* + *mP* itérations. Il est aussi possible, comme le préconisent certains auteurs, notamment Gelman et Rubin (1992, [4]), d'utiliser plusieurs cycles parallèles afin d'assurer l'indépendance.

2.3. Analyse

A ce stade de l'algorithme, on a plusieurs jeux de données complets contenant les valeurs latentes imputées. Avant toute analyse, il faut recréer le jeu de données binaires correspondant par troncature, ce qui est trivial puisque nous avons fixé le seuil à 0 pour toutes les variables. La variable binaire sera donc égale à 0 si la variable latente associée est négative, et sera égale à 1 sinon.

Ensuite, on peut mener les analyses souhaitées séparément sur chacun des jeux de données.

---

[1] La loi de Wishart notée W est la loi suivie par XX',avec X un vecteur gaussien

2.4. Combinaison des résultats

On a obtenu un estimateur de la quantité d'intérêt pour chaque jeu de données imputé, il faut alors combiner ces résultats pour obtenir un estimateur unique dont on souhaite en outre connaître la variance. Soit *Y* notre variable d'intérêt, et *Q* la quantité à estimer (par exemple la moyenne de Y).

Soit $\hat{Q}$ l'estimateur de cette quantité et $U = U\left(Y_{obs}, Y_{manq}\right)$ sa variance, on a alors :

$$\frac{\left(Q - \hat{Q}\right)}{\sqrt{U}} \sim N(0,1)$$

On a *m* estimateurs différents de Q, l'estimateur global s'obtient naturellement par :

$$\hat{Q} = m^{-1} \sum_{m} \hat{Q}^{(l)}$$

Il reste maintenant à déterminer la variance de cet estimateur : elle est composée d'une variance inter-imputations : $B = (m-1)^{-1} \sum_l \left(\hat{Q}^{(l)} - \overline{Q}\right)^2$ et d'une variance intra-imputations $U = m^{-1} \Sigma_l U^{(l)}$.

La variance totale est la somme de ces deux composantes : $T = \left(1 + m^{-1}\right)B + U$

3. Application

3.1. Origines de la méthode : réflexions sur un questionnement en blocs incomplets

Le problème qui a été à l'origine de ces réflexions méthodologiques portait sur la faisabilité d'un questionnement en blocs incomplets sur des données de consommation déclarées par période. Ces données sont tirées de grandes enquêtes, elles ont valeur de données de cadrage et servent à prendre des décisions stratégiques. Pour des raisons de confidentialité, nous ne donnerons pas ici le secteur d'activité concerné, mais ceci ne nous gênera pour le traitement de l'exemple. Actuellement, trois segments du marché sont suivis avec des enquêtes différentes, sur des modes différents, ce qui engendre des chiffres différent pour les marchés suivis dans plusieurs enquêtes, et l'impossibilité d'étudier les croisements (consommation conjointes) de deux produits suivis dans des dispositifs différents.

Plutôt que d'augmenter la longueur de questionnaires déjà roboratifs, l'idée a été de proposer une approche en questionnaires blocs incomplets (c'est-à-dire que tous les individus ne voient pas toutes les questions). On récupère donc des données incomplètes que l'on complète par imputations multiples. La technique de découpage d'un questionnaire en blocs incomplets suivie d'imputations multiples a déjà été étudiée à plusieurs reprises, en particulier par Raghunathan et Grizzle sous le nom de *Split Questionnaire Survey Design*.[11] Ils l'appliquent sur le *Cancer Risk Behavior Survey* Design, une étude américaine sur le mode de vie des ménages, leurs connaissances sur le cancer, et l'impact de celui-ci sur leur mode de vie. Ils comparent ensuite les résultats à ceux issus de l'utilisation du questionnaire entier, et du traitement des seules valeurs disponibles (available cases analysis). Le modèle utilisé dans les imputations est celui de J Schafer, General Location.

Dans notre contexte les avantages pratiques de cette approche sont évidents : une seul mode de collecte, une économie des coûts fixes avec une seule étude, une seule donnée de référence, des échantillons plus importants, une meilleure qualité des réponses associée à des questionnements moins rébarbatifs. Du point de vue du statisticien d'enquête amené à faire de l'imputation, l'assurance d'un arrangement de valeurs manquantes ignorable car construit comme tel au départ.

Du fait de la volonté d'équilibrer la charge entre les différents individus, et les effectifs observés sur les différentes variables, on obtient des arrangements de valeurs manquantes non monotones, donc il faut travailler avec des méthodes multivariées.

L'étude support concernait les données de consommation sur 165 postes pour 40000 individus. Les valeurs moyennes sont très faibles : 60% des indicateurs sont inférieurs à 5%, 40% sont inférieurs à 2.5% …

3.2. Simulations

L'étude de simulation avait comme objectif principal de valider la méthode d'imputation multiple en la comparant à une méthode validée pour le problème. Puisque les données sont exclusivement binaires, nous pouvons indifféremment appliquer le modèle loglinéaire de Schafer, ou la méthode par séquences de régression logistiques. C'est cette dernière solution qui a été retenue avec le logiciel IVEWARE de Raghunathan.

Nous avons repris le principe des simulations décrit dans l'article de Raghunathan et Grizzle : à partir des données originelles, nous avons sélectionné 40 variables et généré 100 échantillons complets en utilisant les valeurs des moyennes et le tableau croisé des variables du fichier initial. Les moyennes initiales sont entre 1.59% et 27.9%. Les croisements les plus faibles sont de l'ordre de 5 ou 6 pour 10000. Nous avons pour cela utilisé la méthode de Berens [6] détaillée au paragraphe 2.1.3. Ensuite, nous avons troué nos échantillons avec des schémas MCAR différents et avec des taux de valeurs manquantes à 5, 10, 15 et 30%, ce qui fait $500 = 100 + (4 \times 100)$ échantillons de 40000 observations. Sur chacune des simulations et pour chaque taux de valeurs manquantes, nous avons lancé des imputations avec 5 réplications avec IVEWARE puis avec notre approche sur variables latentes.

Selon leurs auteurs, la méthode des séquences de régressions d'IVEWARE ne nécessite que peu d'itérations (5 à 10 pour converger), mais nous sommes allés jusqu'à 20 du fait des données sur des très faibles proportions. La procédure met environ 1h45 à tourner.

Pour l'approche sur données latentes, en utilisant 5 séquences de Gibbs parallèles et la statistique R de Gelman Rubin, nous avons fixé le nombre d'itérations de 'burn in' à 200, puis une sauvegarde toutes les 15 itérations. Nous avons utilisé un *a priori* 'ridge' comme le suggère Schafer pour stabiliser les covariances sur ces données très rares. A chaque étape, si les matrices manipulées ne sont pas définies positives, alors on les transforme avec la méthode 'square root'. Pour plus de rapidité les parties les plus souvent appelées dans le programme on été compilées en C. Le programme met entre 45 minutes et 1h20 à tourner selon le nombre de valeurs manquantes : en effet, moins il y a de valeurs manquantes, plus il y a de troncatures définies et plus le programme est long.

Pour chaque échantillon complet répliqué, on calcule les moyennes et les croisements entre toutes les variables deux à deux, ce qui fait 40 estimations pour les moyennes et $(40 \times (40-1))/2 = 780$ estimations pour les croisements. On calcule ensuite les intervalles de confiance à 90% simulés des 820 paramètres d'intérêt, en ne gardant pour chacun de ces paramètres que les 90 valeurs centrales pour déterminer les bornes.

Pour l'étude des biais des générés par les imputations, on calcule pour chaque paramètre, le rapport entre les sommes des estimations sur les 100 fichiers complets et les sommes sur les 100 fichiers troués puis imputés.

Pour l'étude des variations, on calcule pour chaque réplication le rapport entre la proportion d'estimations qui tombent à l'intérieur des IC calculés par simulation et la valeur nominale 90%, puis on édite ces ratios. Ainsi un ratio à 100% sur une des réplications de fichier voudra dire que l'on a fait en moyenne avec la méthode d'imputation 'aussi bien' que le fichier complet : par exemple cela veut dire pour les croisements, que sur 780 estimations $702 = 780 \times 0.9$ auront été dans les bornes et $36 = 40 \times 0.9$ pour les moyennes. Remarquons qu'avec cette mesure, on peut avoir sur une des réplications un ratio supérieur à 100% : 108.3% correspond à 39 moyennes sur 40 dans les limites au lieu des 36 attendues … Au final, c'est donc la moyenne de ces mesures sur les 100 réplications qui donnera la qualité des imputations, avec une valeur nominale à 100%.

Afin de voir le lien éventuel entre ces ratios et le niveau des valeurs estimées, on a aussi fait ces calculs en les limitant aux sous ensembles des valeurs estimées inférieures à 2%, 5%,10%,15%,20% et 30%

3.3. Résultats

Comparées aux valeurs théoriques qui ont servi à les générer, les estimations sur les valeurs complètes sont sans biais et la couverture reconstituée des IC est très proche de la couverture théorique, bien que le nombre de simulations (100) soit assez faible pour l'exercice.

3.3.1. Biais

Pour les moyennes, les biais sont pratiquement nuls sur les deux méthodes, sur toutes les plages de valeurs. Pour les croisements, les deux méthodes ont tendance à les sous-estimer légèrement : ainsi les ratios de sommes sont entre 0.992 à 0.998 pour IVEWARE et 0.994 à 0.998 pour notre méthode. (Les sommes des effectifs dans les croisements hors les diagonales sont à plus de 99% des effectifs initiaux).

3.3.2. Estimations dans les intervalles de confiance : rapports des proportions calculées à la proportion nominale 90%

Pour ce qui est des moyennes, la proportion relative d'estimations dans les intervalles de confiance se situe entre 96.7 et 98% pour les deux méthodes entre les taux de valeurs manquantes de 5 et 30%. Les deux approches sont dans les mêmes plages.

Pour ce qui est des croisements, la méthode normale sur données latentes donne de meilleurs résultats et ce quel que soit le taux de valeurs manquantes testé, on note même un décrochage important pour IVEWARE à partir de 30%.

Rapports des proportions calculées à la proportion nominale : sur 780 croisements, et 100 réplications

Taux de valeurs manquantes	***Méthode IVEWARE***	***Méthode Normale latente***
5%	93.4 - 97.0	93.2 - 97.1
10%	91.7 - 94.3	92.0 - 94.6
15%	87.6 - 92.5	86.9 - 94.0
30%	62.4 - 68.9	79.5 - 86.7

3.3.3. Pourquoi centrer l'analyse sur les croisements ?

Outre le fait qu'il est très facile de vérifier ainsi la qualité des imputations, quand on suit ce type de marché on s'intéresse beaucoup aux consommations conjointes de produits, ceci permet de définir des gammes, et de voir leur couverture en termes de consommateurs touchés. Un grand nombre d'indicateurs sont dérivés de ces croisements : par exemple, le nombre de personnes ayant consommé un, deux, trois produits dans la gamme $\{A, B, C, D, ...\}$, ou le nombre de consommations sur une période donnée, la proportion de consommateurs non touchés par aucun des éléments de la liste, etc …

3.3.4. Discussion : différences entre notre approche et IVEWARE

Les deux approches réalisent théoriquement la même chose sur ces données purement binaires, et elles sont mises en œuvre en parallèle sur les mêmes données : il n'y a donc aucune raison théorique d'avoir des différences entre les deux méthodes.

Une première hypothèse serait à chercher dans la mise en œuvre d'IVEWARE : sur ces données très dispersées, peut être qu'une valeur aussi grande que 20 itérations n'est pas suffisante.

Une autre hypothèse est dans l'implémentation de notre méthode : sachant que le problème portait sur des petites valeurs, nous avons sélectionné les routines de calcul les plus précises. Par exemple, un élément central souvent utilisé dans le tirage des lois tronquées, est le calcul des scores de la loi normale inverse pour des probabilités très proches de 1 ou de 0. Nous avons utilisé le programme en C de P.J. Acklam [16] qui exploite au mieux la précision possible sur la machine qui fait les calculs. Les quantiles de la loi normale bivariée, le tirage dans les lois normales multivariées, la stabilisation des matrices de variance avec la méthode de Higham [5], sont aussi intégrées à partir d'algorithmes

sélectionnés. Les parties les plus intensives en calcul sont écrites en C et compilées pour permettre des centaines d'itérations dans des durées raisonnables.

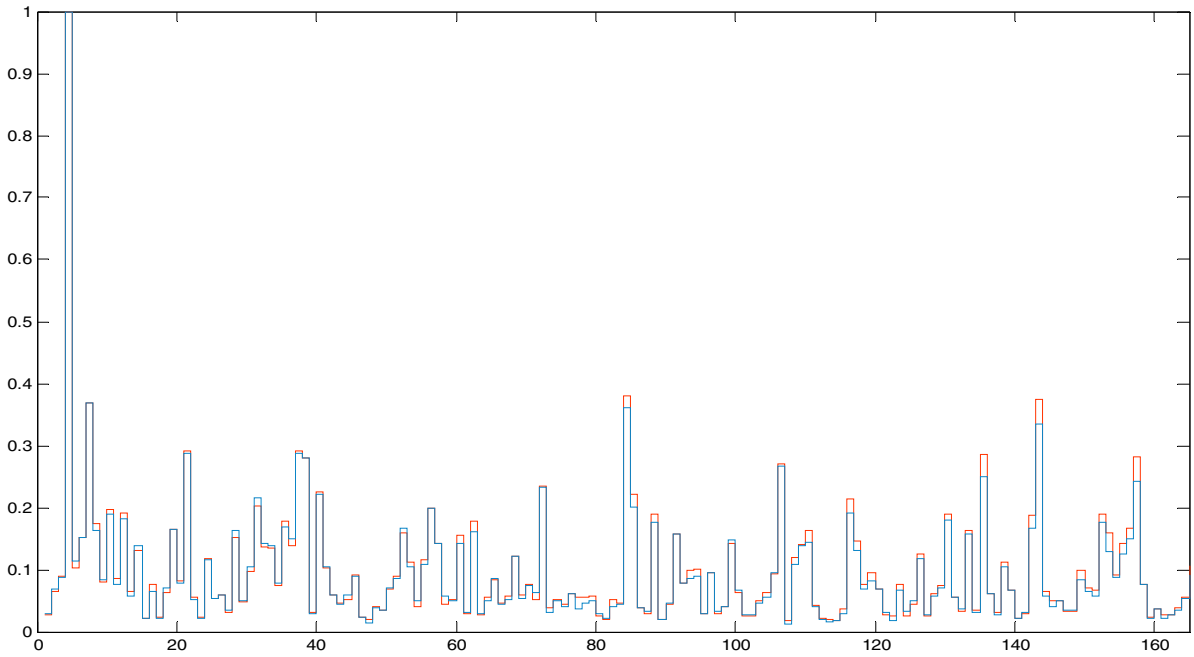
Au final, nous considérons les deux méthodes comme équivalentes du point de vue théorique, mais notre implémentation est meilleure du point de vue de la précision.

3.4. Quelques résultats sur les données complètes (165 variables)

Nous n'avons pas répété les simulations sur les données complètes, car il faut alors environ 2 à 3 minutes par itération. Les moyennes initiales sont maintenant entre 0.079% et 27.9%, Sur les essais réalisés notre programme se comporte mieux que IVEWARE, les imputations sont très bonnes, mais on note une légère sous-estimation des interactions (effectifs croisés) les plus fortes, et une légère surestimation des interactions les plus faibles. Ceci est typiquement la conséquence des transformations faites pour régulariser les matrices de corrélations qui résument les interactions : on a tendance à les 'lisser' un peu.

Un exemple est dans le graphique suivant, qui montre les proportions de consommateurs de chacun des 165 produits parmi ceux qui consomment le produit 4. La ligne rouge (données originelles) est presque toujours au dessus de la bleue (données imputées) pour des croisements importants, et en dessous sinon.

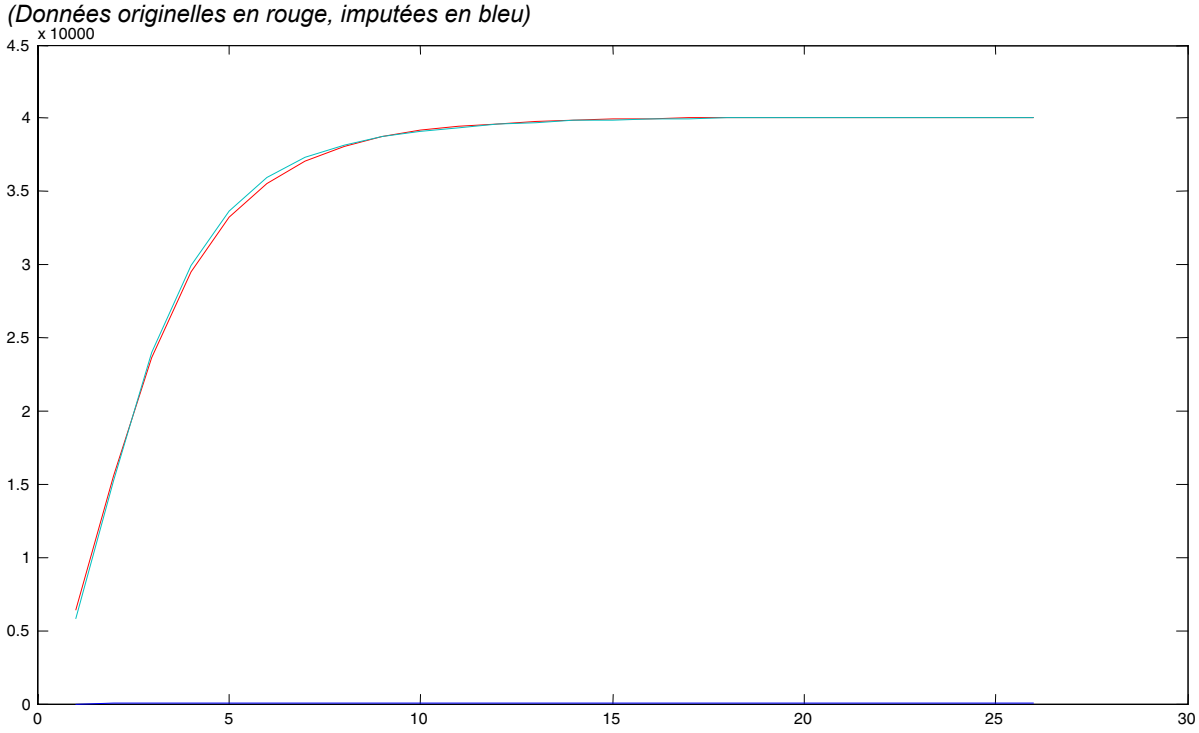
Fig.1 : Proportion de consommateurs de chacun des 165 produits parmi les consommateurs du produit 4.
(Données originelles en rouge, imputées en bleu)



Sur ces indicateurs reconstruits, la précision des estimations est importante : il faut par exemple que la proportion de consommateurs touchés par aucun des produits d'une gamme soit à 2 ou 3% près celle des données originelles ; sur ce type de calculs un biais même minime sur les croisements peut avoir des conséquences.

Voici un autre exemple de graphique de comparaison entre les données originelles et imputées: la distribution cumulée, c'est-à-dire le nombre de personnes qui consomment au moins x produits.

Fig. 2 : Complémentaire de la distribution cumulée : nombre d'individus consommant moins de x produits parmi K (K= allant de 1 à 40, nombre d'individus donnés en dizaines de milliers)
(Données originelles en rouge, imputées en bleu)



4. Conclusion

Les résultats de ce test ont permis de valider la méthode, et de disposer d'un outil beaucoup plus précis que les autres algorithmes que nous avons testé. Les programmes de Schafer CAT, MIX et PAN comportent beaucoup de bugs et ne sont plus maintenus, IVEWARE ou le portage de FCS dans SPSS 17 n'ont pas donné de bons résultats. Cette précision était nécessaire vu les tailles de fichiers manipulés et la rareté des évènements suivis.

Même si l'algorithme parait complexe à énoncer, il est très simple à programmer : hormis le tirage et le calcul des probabilités inverses de la loi normale, tous les autres composants du programme sont des routines d'algèbre linéaire très simples et largement disponibles. Par rapport aux programmes disponibles, nous avons totalement la main dessus et la donc la possibilité d'éditer de nombreux diagnostics de convergence en isolant des sorties (statistique R de Gelman Rubin, graphiques d'autocorrelation, etc.…).

Si l'on doit replacer cette approche parmi les outils disponibles pour faire de l'imputation multiple, on dira qu'étant limitée aux variables binaires, elle est donc d'application très spécialisée par rapport à certaines autres routines plus générales. Elle n'en constitue pas moins une extension intéressante à la palette de modèles déjà proposés par Schafer. Elle permet en effet de pousser plus loin le modèle normal qui est le seul qui marche vraiment sur des données réelles de très grande dimension. C'est aussi et surtout une façon plus correcte de traiter les variables catégorielles que les méthodes basées sur des arrondis de valeurs continues [1, 2, 3, 14] qui sont toutes génératrices de biais à plus ou moins grande échelle [8].

Bibliographie

[1] Ake CF. *Rounding after multiple imputation with non-binary categorical covariates* SUGI 30 Proceedings 2005, 112–30, pp. 1–11.

[2] Allison PD. *Imputation of categorical variables with PROC MI* . SUGI 30 Proceedings 2005, 113–30, pp. 1–14.

[3] Coen A. Bernaards, Thomas R. Belin, Joseph L. Schafer : *Robustness of a multivariate normal approximation for imputation of incomplete binary* data Statistics in Medicine Volume 26 Issue 6, Pages 1368 - 1382

[4] Gelman, A. and Rubin, D. B. (1992), *Inference from Iterative Simulation Using Multiple Sequences* Statist. Sci. Volume 7, Number 4 (1992), 457-472

[5] Higham, N.J. [2000*], Computing the nearest correlation matrix — A problem from finance*, Department of Mathematics, University of Manchester, Numerical Analysis Report, 369 (available from http://www.ma.man.ac.uk/~higham/)

[6] *Macke, J. H., P. Berens, A. S. Ecker, A. S. Tolias and M. Bethge:* Generating Spike Trains with Specified Correlation Coefficients. Neural Computation, 1-27 (in press) (02 2009)

[7] Horton , Nicholas J. and Lipsitz , Stuart R.(2001) *Multiple imputation in practice: comparison of software packages for regression models with missing variables*, Statistical computing software reviews, The American Statistician,August 2001, Vol. 55, N3.

[8] Horton, N.J., Lipsitz, S.R., and Parzen, M. (2003), *A Potential for Bias When Rounding in Multiple Imputation*, American Statistician, 57, 229-232

[10] Roderick J. A. Little and Donald B. Rubin (1987), *Statistical analysis with missing data*, 2nd edition (New York: Wiley).

[11] Raghunathan, Trivellore E. and Grizzle, James E. (1995), *A spilt questionnaire survey design*, Journal Of the American Statistical Association, Vol.90, N°429, March 1995, 54-63.

[12] Raghunathan, Trivellore E. , Lepkowski, James M. , VanHoewyk, John and Solenberger, Peter, *A multivariate technique for multiply imputing missing values using a sequence of regression models*, n 081, Survey Methodology Program, Institute for Social Research of the University of Michigan.

[13] Raghunathan, T. E., Solenberger, Peter W., Van Hoewyk, John, *IVEware: Imputation and Variance Estimation Software* http://www.isr.umich.edu/src/smp/ive/

[14] Recai M Yucel, Yulei He, Alan M Zaslavsky. *Using Calibration to Improve Rounding in Imputation* The American Statistician

[15] Rubin, Donald B. (1987), *Multiple imputation for nonresponse in surveys*, New York: John Wiley and Sons.

[16] Sprouse C, Acklam PJ. *An algorithm for computing the inverse normal cumulative distribution function* http://home.online.no/~pjacklam/notes/invnorm/index.html

[17] Schafer, J.L. (1997), *Analysis of Incomplete Multivariate Data*, London: Chapman and Hall.

[18] Schafer, J.L. – page for Software for multiple imputation
http://www.stat.psu.edu/~jls/misoftwa.html

[19] Shum Matthew, GHK simulator: get draws from truncated multivariate normal distribution http://www.econ.jhu.edu/people/shum/e2901/ghk_desc.pdf

[20] Van Buuren, Stef *Multiple imputation of discrete and continuous data by fully conditional specification* Statistical Methods in Medical Research, Vol. 16, No. 3, 219-242 (2007)