Imputation Multiple de données catégorielles

Une approche basée sur un modèle Normal latent

Index

1. Introduction - Contexte
2. Imputations Multiples - le cadre
3. Algorithme
4. Simulations - Résultats
5. Conclusion

Imputation Multiple de données catégorielles

Une approche basée sur un modèle Normal latent

i

Authors TNS-SOFRES - ENSAE
Philippe Périé - Anne de Moliner

Created
16 mars 2009

Introduction - Contexte

i

Imputation Multiple de données catégorielles

Origine de la réflexion

- Faisabilité d'un ***questionnement en blocs incomplets*** sur des données de consommation déclarées par période.
- Actuellement, du fait de l'émiettement de ce marché, trois segments sont suivis avec des enquêtes différentes, sur des modes différents :
  - Des chiffres différents pour les marchés suivis dans plusieurs enquêtes,
  - 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.*

Introduction - Contexte

- Cette réflexion nous a amené à chercher des solutions pour le traitement des données catégorielles (en particulier binaires) de grande dimensionnalité, portant sur des évènements rares. De telles données posent pas mal de problèmes au praticien du fait des effectifs et donc des croisements très faibles.
- 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. …

- Pour valider, nous avons 'troué' un fichier complet pour simuler des valeurs manquantes, que nous avons complétées ensuite par imputation multiple

- Après pas mal de tests avec les méthodes existantes, nous avons développé une nouvelle méthode qui en est une extension, et qui a donnée de meilleurs résultats en particulier pour les cellules à faibles effectifs. C'est cette méthode que nous présentons

Imputations Multiples - le cadre

i

Imputation Multiple de données catégorielles

Principe

- **Idée:** remplacer chaque valeur manquante par 2 ou plus valeurs acceptables représentant une distribution de possibles (Rubin, 1987).
- On obtient *m* fichier complets, que l'on peut analyser avec des méthodes classique, les estimations sont la moyenne des estimations dans chaque fichier

- **Avantages :**
  - *Possibilité d'utiliser des méthodes classiques d'analyse*
  - *Permet d'incorporer l'incertitude sur la Non Réponse*
  - *Améliore l'éfficacité des estimations*
  - *Permet une inférence valide (variances) sous certains modèles de non réponse*

Principe

- Soit Y un vecteur de variables d'intérêt incomplètes de dimension k. Les parts observées et manquantes de Y sont notées Y_{obs} et Y_{manq} respectivement.
- Soit R le vecteur des indicatrices de réponse dont les composantes valent 1 si Y est observé et 0 sinon. On définit aussi X un ensemble de covariables complètement observées sur les mêmes sujets.

- **Approche bayesienne** (Rubin, 1987): tirer des imputations multiples pour simuler la distribution Bayesienne a posteriori : $P(\theta / Y_{obs}, R)$

  Complexe, plus simple à faire quand on *augmente* les donnes observées des données manquantes : $P(\theta / Y_{obs}, Y_{manq}, R)$

- Les modèles d'imputation reposent sur l'hypothèse que le **mécanisme de non-réponse est ignorable** , c'est-à-dire que $P(Y|X, R = 0) = P(Y|X, R = 1)$.

Principe - Hypothèses

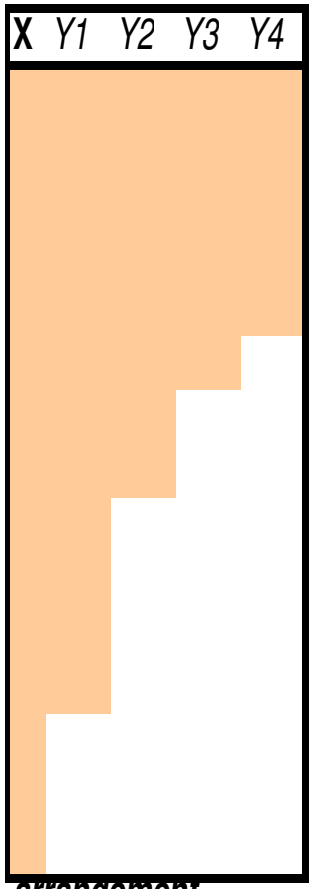
- Soit Y un vecteur de variables d'intérêt incomplètes de dimension k. Les parts observées et manquantes de Y sont notées Yobs et Ymanq respectivement.
- Soit R le vecteur des indicatrices de réponse dont les composantes valent 1 si Y est observé et 0 sinon. On définit aussi X un ensemble de covariables complètement observées sur les mêmes sujets.
- **Approche bayesienne** (Rubin, 1987): tirer des imputations multiples pour simuler la distribution Bayesienne a posteriori : P(
  - Complexe, plus simple à faire quand on *augmente* les donnes observées des données manquantes :
- Les modèles d'imputation reposent sur l'hypothèse que le mécanisme de **non-réponse est ignorable** , c'est-à-dire que P(Y|X, R = 0) = P(Y|X, R = 1).

Arrangement des données manquantes

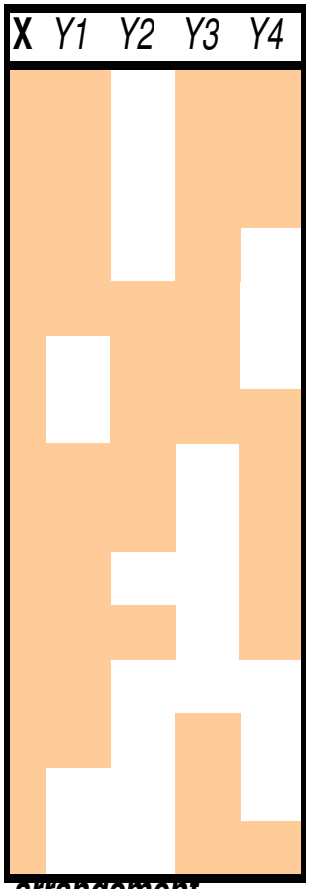
- Arrangement **monotone** :
  - des imputations multivariées peuvent être obtenues par une factorisation de tirages dans des lois univariées

    (Y1/X puis Y2/Y1,X puis Y3/Y2*,Y1*,X etc.)*

- Arrangement **non monotone** :
  - Il faut travailler avec des méthodes d'imputation multivariées.
  - La complexité des lois manipulées interdit alors les estimations par calcul analytique, et l'on a recours à des méthodes itératives.
  - **Gibbs Sampling** : on tire dans une loi jointe complexe par tirages successifs dans des lois conditionnelles plus simples.



arrangement monotone



arrangement général

Deux classes de méthodes

- 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. **Joint Modelling** est disponible dans SAS pour le modèle normal avec la procédure MI, ou avec le logiciel NORM de Schafer.
2. **FCS** (ou **séquences de régressions**) est disponible dans IVEWARE depuis SAS, MICE sur R, Multiple Imputations sur SPSS 17 …

Joint Modelling (Rubin 1987; Schafer 1997)

- On impose une densité de probabilité sur les données complètes (Y_{obs},Y_{manq},θ) **ET** sur le mécanisme de non-réponse (par exemple, normal or loglinéaire)
- On crée les imputations avec un processus Bayesien en 2 étapes :
  - Spécifier des distributions a priori des paramètres θ et tirer dans la distribution conditionnelle sachant les données
  - Simuler *m* tirages indépendant de la distribution conditionnelle des données manquantes sachant les observées
- Pour cela on itère

  A l'étape (r) :

  1/ le I-step tire $Y_{manq}^{(r+1)}$ dans $p\left(Y_{manq} \middle/ Y_{obs}, \theta^{(r)}\right)$

  2/ le P-step tire $\theta^{(r+1)}$dans $p\left(\theta \middle/ Y_{obs}, Y_{manq}^{(r)}\right)$

- Les deux étapes sont itérées assez longtemps pour que les estimations se stabilisent, tirer dans : $P\left(Y_{manq}, \theta \middle/ Y_{obs}\right)$

Séquences de régression (Raghunathan), FCS (Van Bureen)

- Faire l'économie de la spécification explicite d'une loi jointe complexe,
- Itérer les tirages dans les lois conditionnelles, une pour chaque variable traitée , on cherche à obtenir la loi jointe.
- Avantages :
  - Grande flexibilité pour des données d'enquête dans lesquelles on peut avoir des dépendances complexes (censures, croisements vides, …)
- Inconvénients :
  - Propriétés statistiques 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.

Algorithme

i

Imputation Multiple de données catégorielles

Notre méthode : Imputation sur variables latentes normales

- Cadre normal multivarié de Schafer (1997)
- On spécifie un modèle latent, dont les troncatures sont déterminées par les valeurs observées. Les valeurs imputées sont donc tirées conditionnellement à ces troncatures.

- **Algorithme :**

1. **Initialisation** : déterminer les vecteurs normaux latents associés aux valeurs observées binaires
2. **Méthode MCMC: Gibbs sampling,** alterner jusqu'à convergence les 2 phases :
   - I-Step *(spécifique du fait de la coexistence des variables binaires et latentes)*
   - P-Step
3. **Imputations :** une fois la convergence obtenue, tirages pour imputations multiples.
4. **Retour aux variables binaires** : Appliquer les seuils aux vecteurs normaux latents

Les variables latentes

- Aux variables binaires Y_i, correspondent des variables latentes Y_i^* normales qui les déterminent.

Y_i binaire	Y^*_i continue
$Y_i \sim B(p)$	$Y^* \sim N(\mu,\Sigma)$

$$Y_i = 0 \quad \Leftrightarrow \quad Y^*_i > S_i$$

- S est le vecteur des seuils que l'on peut fixer à 0 sans perte de généralité

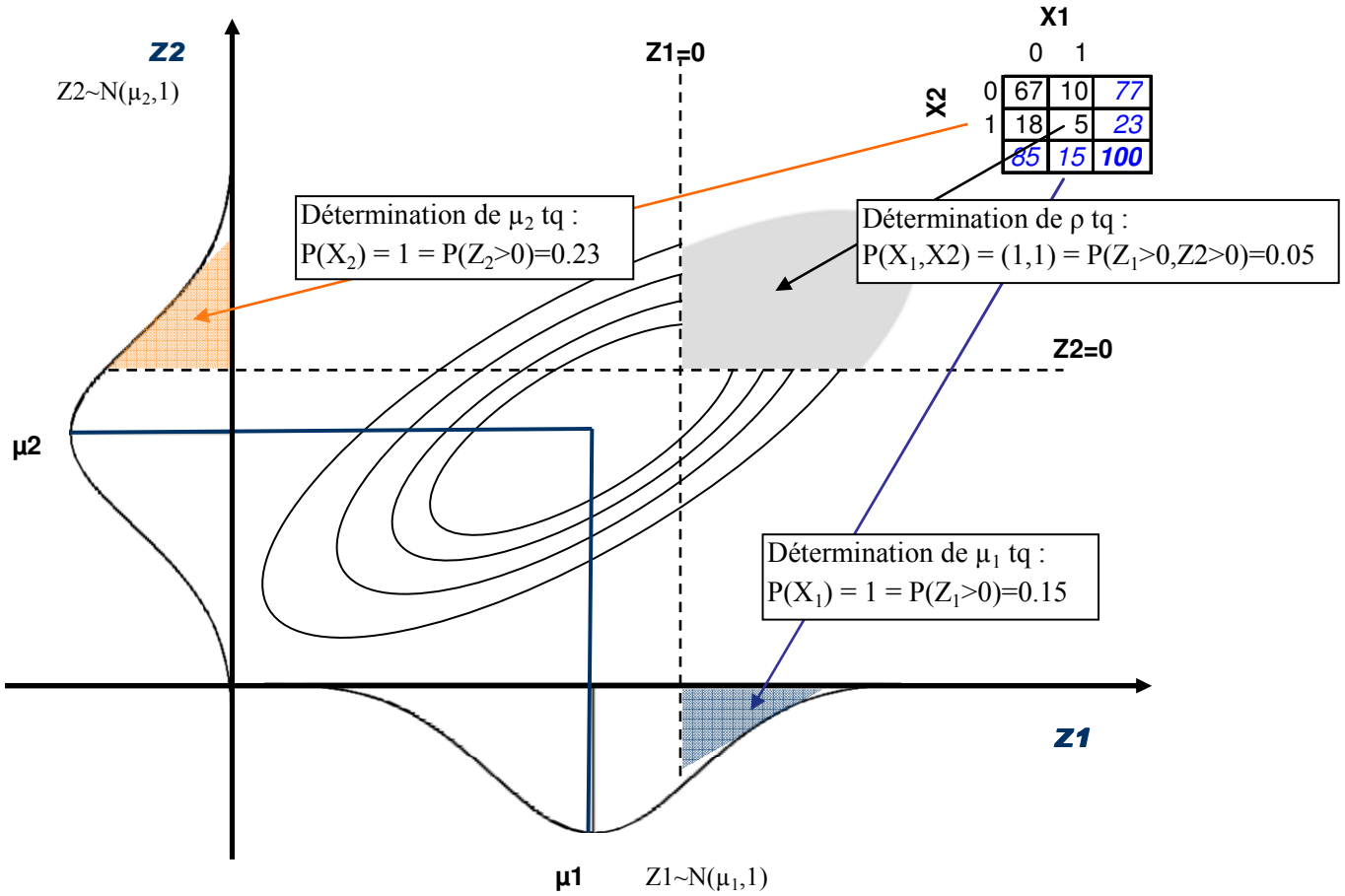
Pourquoi travailler sur des variables latentes normales ?

- D'abord, une spécification qui a du sens : fonction latente d'utilité
- Des limites logicielles : variables binaires de grande dimension, avec de très faibles occurrences
- Retour dans le cadre normal: plus simples à programmer, bibliothèque de routines la plus large et optimisation numérique dans ces cas extrêmes.

Illustration : dans le cas de deux variables

D'un tableau croisé 2x2 à la loi Normale latente associée

- Obtenir μ_1, μ_2, ρ paramètres de la loi normale bivariée associée au tableau croisé



Méthode de Berens dans le cas n >2

- Les moments latents (r, Σ) découlent des moments binaires observés (γ,Λ):
  - $r_i=\Phi(\gamma_i)$
  - $\Sigma_{i,i}=\Phi(\gamma_i)\ \Phi(-\gamma_i)$
  - $\Sigma_{i,j}=\Psi(\gamma_i, \gamma_j,\Lambda_{i,j})$

  - Φ est la fonction de répartition d'une loi normale
  - Ψ la probabilité pour que deux variables normales de corrélation $\Lambda_{i,j}$ soient supérieures à leurs seuils respectifs γ_i et γ_j

Méthode de Berens dans le cas n >2

- On tire le vecteur latent conditionnellement au vecteur binaire (observé): il suit une loi normale tronquée multivariée, de densité :

$$fr(x)=C \times (2\pi)^{-N/2} |\Sigma|^{-N/2} \exp\{(x-\mu)\,\Sigma^{-1}\,(x-\mu)'\}\, IR(x)$$

  - Ir(x) = 1 ⇔ les contraintes imposées par les valeurs binaires sont respectées
  - C est une constante de normalisation

Tirage dans des lois Normales tronquées : Méthode GHK

(Geweke, Hajivassiliou et Keane)

- On veut tirer le vecteur *Y* selon une loi normale tronquée avec les restrictions (A,B) :

$$Y \sim TN(\gamma, \Lambda, A, B) \Leftrightarrow Y \sim N(\gamma, \Lambda)\, s.c.\ A < Y < B$$

- Soit $\Lambda^{1/2}$ la décomposition (inférieure) de Choleski de Λ, alors

$$Y = \gamma + \Lambda^{1/2} U \text{ et } Y \sim N(\gamma, \Lambda)\, s.c.\ A < Y < B$$

avec U un vecteur normal centré réduit

- Tirer dans une loi tronquée multivariée = tirer séquentiellement dans des lois tronquées univariées

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

$$\Lambda^{1/2} = \begin{pmatrix} l_{1,1} & 0 & 0 & 0 \\ l_{2,1} & l_{2,2} & 0 & 0 \\ \dots & \dots & l_{i,i} & 0 \\ l_{p,1} & l_{p,2} & \dots & l_{p,p} \end{pmatrix}$$

Intégration des covariables : analyses factorielles

- Les covariables X, entièrement observées, sont intégrées dans le modèle par le biais de leurs facteurs issus d'analyses factorielles.
- Facteurs= combinaisons de variables orthogonales entre elles et centrées, considérées comme normalement distribuées.
- Modification de la phase de tirage : les facteurs conditionnent les valeurs latentes tirées.

Une nouvelle décomposition de Choleski de la matrice de variance covariance:

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

Phase d'imputation : I-Step

- *I-Step légèrement différente de celle du modèle normal de Schaffer*:

  Y_{obs} binaires,

  => Il faut tirer les valeurs latentes Y^*_{obs} conditionnellement à ces Y_{obs} qui déterminent des troncatures. $\left(Y^*_{obs} / Y^*_{manq}, Y_{obs}, X, \mu, \Sigma\right) \sim TN\left(\mu_{cond}, \Sigma_{cond}, S\right)$

Ensuite seulement, on peut tirer Y^*_{manq} sachant Y^*_{obs} …

- Ainsi à chaque itération :

1. On tire d'abord les valeurs latentes dans la loi normale tronquée conditionnelle
2. On tire ensuite les valeurs latentes des valeurs manquantes conditionnellement aux paramètres et aux valeurs latentes des variables observées.

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

Phase P Step :

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

1. On commence par tirer Σ dans une loi de Wishart inverse

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

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

2. On tire ensuite μ dans une loi normale

$$\left(\mu/\Sigma, X, Y_{obs}, Y_{manq}\right) \sim N\left(\overline{Y}, N^{-1}\Sigma\right)$$

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.

Utilisation d'un a priori informatif

- Avec un *a priori* non informatif (de Jeffreys) pour les variances: convergence très lente vers la loi de la distribution jointe et sous-estimation des corrélations

- On spécifie une loi *a priori* de la matrice des corrélations pour stabiliser l'inférence: a priori informatif

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

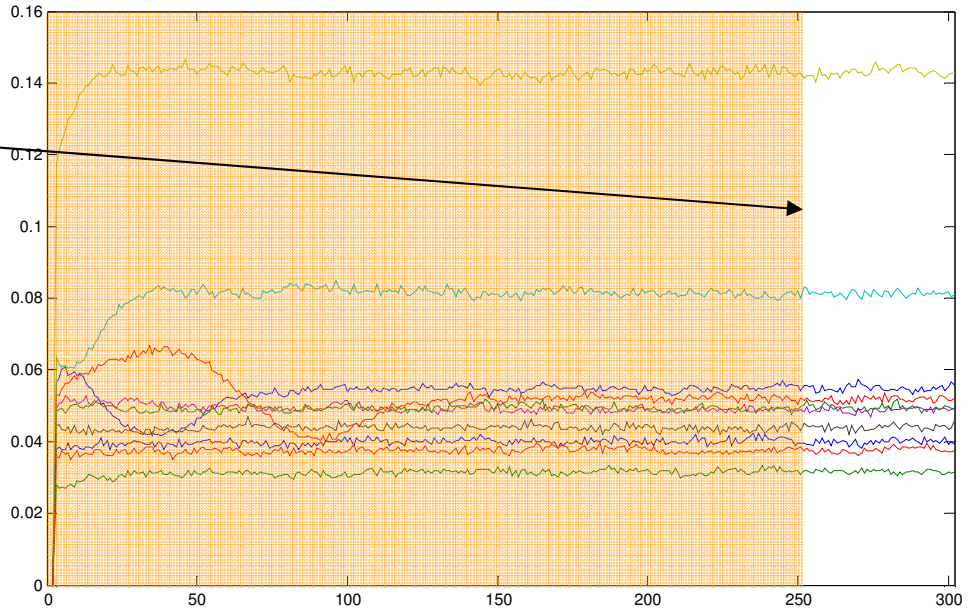
- Par exemple, on peut prendre:
  - L la corrélation obtenue sur une enquête précédente
  - m: poids de l'*a priori* face aux observations.

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 (*phase d'initialisation ou burn in)*
- Ensuite: on stocke un tirage toutes les *P* itérations, avec *P* suffisamment grand (usuellement entre 100 et 1000). Les tirages peuvent être considérés comme indépendants s'ils sont suffisamment espacés.

Exemple pour 10 paramètres, la partie 'Burn In' est en orangé

Convergence au bout de 250 itérations



Analyse

- A ce stade de l'algorithme, on a plusieurs jeux de données complets contenant les valeurs latentes imputées.
- On recrée le jeu de données binaires correspondant par troncature, ce qui est immédiat 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.
- Analyses séparées sur chacun des jeux de données.

Simulations - Résultats

i

Imputation Multiple de données catégorielles

Simulations

- Tester la méthode d'imputation multiple en la comparant à une méthode validée pour le problème.
- **Données exclusivement binaires** : le modèle loglinéaire de Schafer, ou séquences de régression logistiques. C'est cette dernière solution qui a été retenue avec le logiciel IVEWARE de Raghunathan.
- **Principe des simulations :**
  - A partir des estimations de θ sur les données complètes, déterminer θ* du modèle normal latent, générer 100 échantillons par tirage dans la loi, trouer les échantillons selon un schéma MCAR à 5,10,15,30% de manquants : donc 500 fichiers.
  - Nous avons pris un extrait de fichier avec 40 variables binaires, ce qui fait entre les moyennes et les croisements 40+(40x39/2) = 820 paramètres.
  - En prenant les 90 estimations centrales des 100 fichiers complets sur les 820 paramètres on détermine des IC simulés à 90%
  - On regarde pour tous les taux de valeurs manquantes, le proportions d'estimations qui tombent dans ces IC et on les rapporte à 90%.
  - Au final 100% sur un paramètre (ou un ensemble de paramètres) veut dire que l'on a fait 'aussi bien' que le fichier complet
  - On peut aussi calculer les biais

Résultats

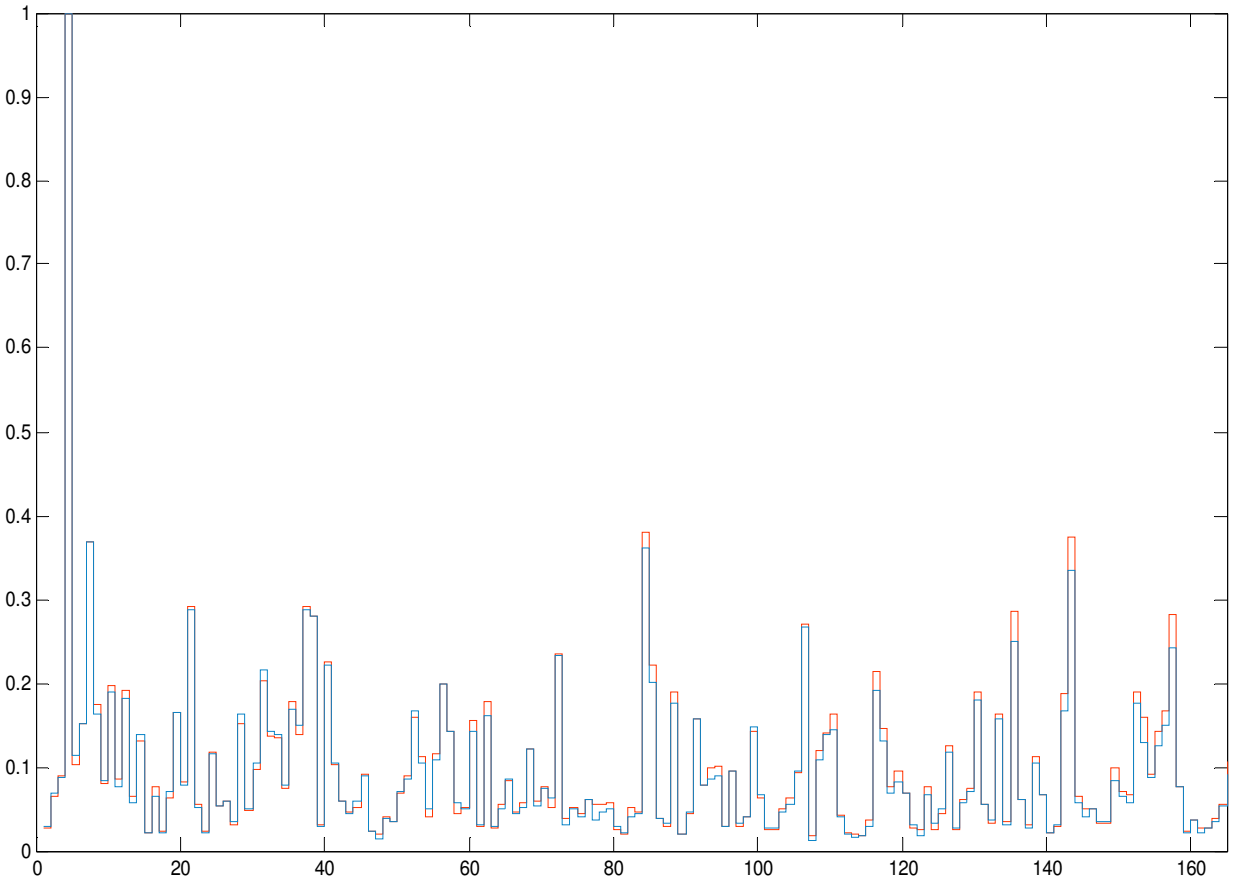
- **Biais :**
  - Pour les moyennes, les biais sont pratiquement nuls sur les deux méthodes, sur toutes les plages de valeurs. Pour les croisements, les deux méthodes ont tendance à les sous estimer légèrement : ainsi les ratios sont entre 0.992 à 0.998 pour IVEWARE et 0.994 à 0.998 pour notre méthode.
- **Couverture des IC :**
  - Pour ce qui est des moyennes, on est 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%.

Proportions d'estimations des effectifs croisées dans les IC : 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

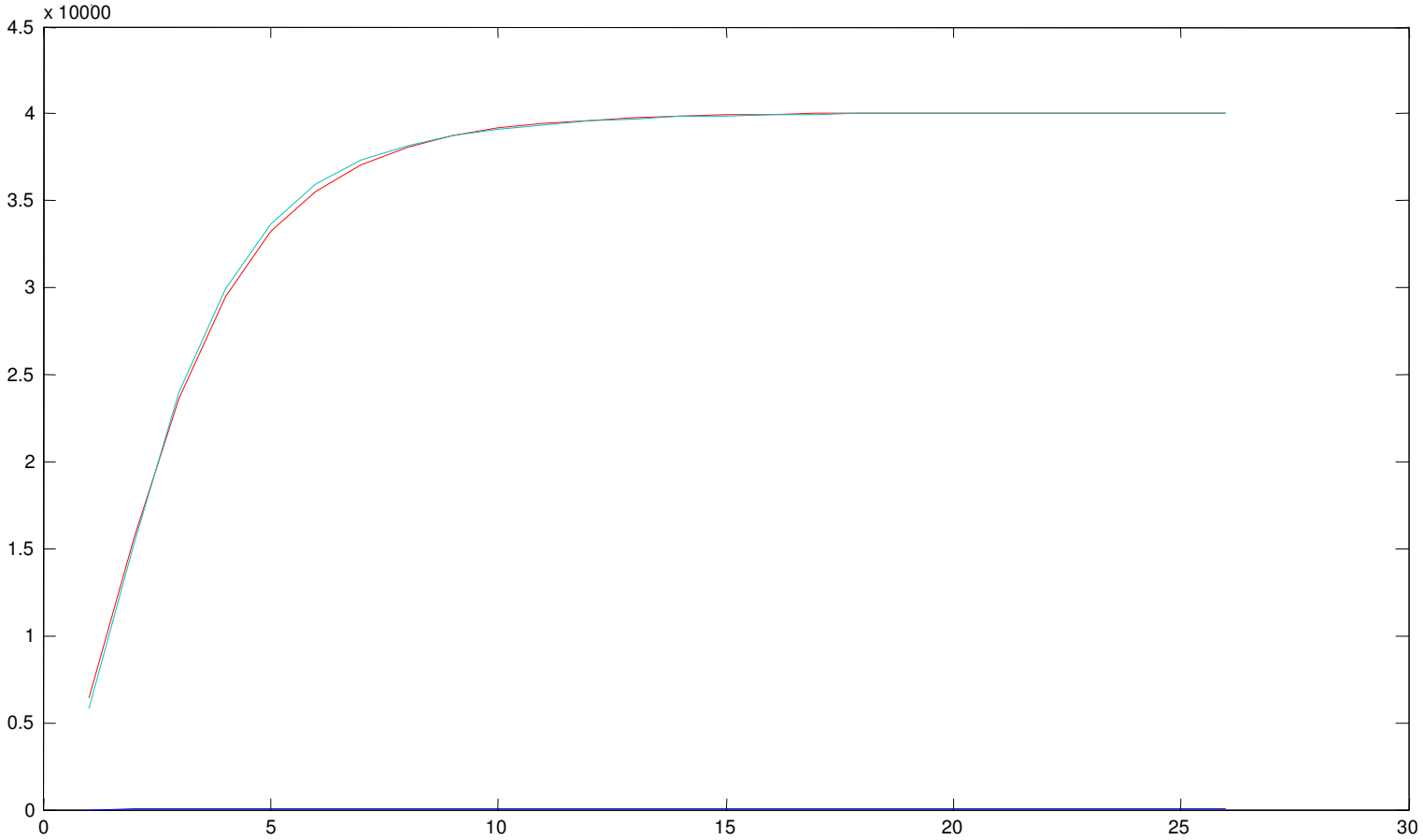
Reconstitution des croisements

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



Reconstitution des distribution cumulées

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)



Conclusion – Comparaison avec 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.
- 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.

Conclusion

i

Imputation Multiple de données catégorielles

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. 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.

Conclusion

- 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 qui sont toutes génératrices de biais à plus ou moins grande échelle.

Merci