Le traitement des donnees manquantes dans le logiciel SICLA

Gilles Celeux

▶ **To cite this version:**

Gilles Celeux. Le traitement des donnees manquantes dans le logiciel SICLA. RT-0102, INRIA. 1988, pp.15. inria-00070064

HAL Id: inria-00070064
https://inria.hal.science/inria-00070064

Submitted on 19 May 2006

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

INRIA

UNITÉ DE RECHERCHE
INRIA-ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
B.P.105
78153 Le Chesnay Cedex
France
Tél.(1)39.63.55.11

Rapports Techniques

N° 102

Programme 5

LE TRAITEMENT DES DONNÉES MANQUANTES DANS LE LOGICIEL SICLA

Gilles CÉLEUX

Décembre 1988

LE TRAITEMENT DES DONNEES MANQUANTES DANS LE LOGICIEL SICLA

Gilles Celeux
INRIA
Domaine de Voluceau, Rocquencourt, B.P. 105
78153 LE CHESNAY Cedex - FRANCE

RESUME

On présente les techniques du logiciel SICLA d'attribution de valeurs à des données manquantes au hasard. Ces techniques relèvent du maximum de vraisemblance à partir de modèles statistiques classiques et certaines utilisent un principe d'attribution aléatoire. Ce rapport contient une typologie des différents types de données manquantes et les caractéristiques des stratégies utilisées dans SICLA pour les données quantitatives et pour les données qualitatives sont commentées.

Mots-clés : données manquantes au hasard, attribution de valeurs, maximum de vraisemblance, principe d'attribution aléatoire.

FILLING IN FOR MISSING DATA IN THE SICLA SOFTWARE

ABSTRACT

We present the imputation procedures for missing data at random available in the SICLA software. These procedures are based on maximum likelihood derived from classical statistical models, and some of them used a random imputation principle. This report reviews the different missing data mechanisms and the features of the SICLA procedures for quantitative and qualitative data are discussed.

Keywords: missing data at random, filling in, maximum likelihood, random imputation principle.

PAPIER RÉCUPÉRÉ ET RECYCLÉ

1. Introduction

La présence de données manquantes est source d'embarras pour les statisticiens et pour les concepteurs de logiciels. Ce rapport est destiné à présenter les solution adoptées pour la prise en compte de données manquantes dans le logiciel interactif de classification automatique SICLA développé à l'INRIA (SICLA (1988)). Afin de bien clarifier la portée de nos choix, leur exposé est précédé de deux paragraphes qui précisent les différentes situations d'occurence de données manquantes et présentent les stratégies classiques de prise en compte de données manquantes fortuitement. Cette partie préliminaire privilégie le point de vue du concepteur de logiciel qui doit proposer des solutions suffisamment générales pour s'adapter à de nombreuses et diverses méthodes d'analyse statistique.

2. Nature des données manquantes

Lorsqu'on se trouve confronté à des données manquantes, la première question à se poser est : leur apparition est-elle due au hasard ou, au contraire, à des raisons déterministes ? La réponse à cette question oriente de manière fondamentale la prise en compte des données manquantes. Nous allons donc préciser les termes de cette question en introduisant les définitions de Little et Rubin (1987).

Données Manquantes au Hasard (DMH)

On dira qu'une donnée manque au hasard si l'occurence de son absence est indépendante de la valeur qu'elle prend.

Autrement dit, les données manquantes pour une même variable peuvent être considérées comme un sous-échantillon aléatoire de l'échantillon initial.

Données Manquantes Complétement au Hasard (DMCH)

On dira qu'une donnée manque complétement au hasard si l'occurence de son absence est indépendante de toutes les valeurs que prend l'individu qui présente cette donnée manquante.

Autrement dit, l'ensemble des données manquantes complétement au hasard constitue un sous-échantillon aléatoire des valeurs prises par l'échantillon initial. Par contre, l'ensemble de données manquantes au hasard ne constitue pas nécessairement un sous échantillon aléatoire des valeurs initiales.

Exemple : On considère une population décrite par deux variables, l'âge et le revenu. Si la non connaissance d'un revenu n'est pas liée à sa valeur, il s'agit d'une donnée manquante de type DMH ; si, de plus, elle n'est pas liée à l'âge de la personne interrogée, il s'agit d'une donnée manquante de type DMCH.

Si les données manquantes ne sont pas de type DMH, leur prise en compte doit s'appuyer sur les raisons de leur absence. Si les données manquantes sont de type DMH, leur prise en compte ignore les raisons de leur absence et doit s'appuyer sur les relations entre les variables. Si les données manquantes sont de type DMCH, leur prise en compte ignore les raisons de leur absence et peut ignorer les relations entre les variables.

Pour des variables issues de mesures ou d'observations physiques il est en général assez simple de décider qu'une données manquante est de type DMH ou non. La plupart du temps les données manquantes qui ne sont pas de type DMH sont des données censurées à droite ou à gauche. Une donnée est censurée à droite (resp. à gauche) si on sait seulement qu'elle est plus grande (resp. plus petite) qu'une certaine valeur. Ce type de données se rencontre en contrôle non destructif de qualité, en analyse des données médicales de survie ou lorsqu'une mesure dépasse les capacités d'une machine. De nombreux ouvrages sont exclusivement consacrés à l'analyse statistique à partir des données censurées. Citons, par exemple, Kalbfleish, Prentice (1980) et l'ouvrage en français de Carbon, Gourieroux, Huber et Lecoutre (1989), émanation d'un cours de l'ASU sur l'analyse des données de durée de vie, à paraitre aux éditions Economica. Signalons que l'analyse à partir des données imprécises (données dont on ne connait pas la valeur exacte mais dont on sait qu'elles tombent dans un intervalle) se fait par le même type de techniques utilisées pour des données censurées.

Pour des variables issues d'enquêtes il peut être délicat de déterminer si les données sont de type DMH et encore plus de type DMCH. Pour s'en convaincre, il suffit de considérer l'exemple du revenu et de l'âge. Une non réponse pour le revenu a-t-elle pour cause une omission, ou une volonté de le tenir secret ? Dans ce dernier cas, est-ce simplement un réflexe culturel lié à l'âge de la personne interrogée ou est-ce que la non réponse est due à l'importance du revenu ? En règle générale, il est erroné de considérer que les non réponses d'une enquête sont de type DMH, si la raison de l'absence est douteuse. L'expérience montre, au contraire, qu'elles peuvent apporter beaucoup d'information si on les considère comme des modalités de réponses au même titre que les autres (cf. Lebart, Morineau, Tabard (1977), van der Heijden, Escofier (1988)). On trouvera des techniques spécifiques de prise en compte de non réponses dans les enquêtes dans Rubin (1987) et van der Heijden, Escofier (1988). Ces derniers auteurs examinent l'influence de diverses stratégies de prise en compte de données manquantes sur les résultats de l'analyse des correspondances multiples, méthode centrale dans l'analyse des questionnaires. Par ailleurs, ils caractérisent un type important de données manquantes pouvant intervenir dans les enquêtes: les données non catégorisables qui correspondent à une situation où les modalités de réponses proposées ne permettent pas de décrire certains individus. Pour ce type de données manquantes, ils recommandent de

considérer la non réponse comme une modalité à part entière (que l'on pourrait intituler "autre réponse").

3. Stratégies de prise en compte de données manquantes au hasard

Deux attitudes ont cours pour la prise en compte des données manquantes au hasard. La première part de la considération d'une tâche statistique précise à réaliser (une régression par exemple) et conduit à adapter les calculs pour une prise en compte intelligente des données manquantes. La deuxième pose le problème en termes d'attribution de valeurs aux données manquantes (on parle aussi de reconstitution de données manquantes) ce qui permet ensuite d'utiliser toute technique statistique sans modification.

Le premier point de vue d'adaptation des techniques est peut-être plus souhaitable mais nécessite un gros travail d'analyse et de programmation qui s'avère trop lourd pour la réalisation de logiciels comportant (comme SICLA) de nombreuses et très diverses méthodes statistiques. On trouvera des exemples de calculs statistiques suivant ce point de vue dans Little, Rubin (1987) pour l'analyse de variance et l'analyse de covariance, en particulier, et dans Der Megreditchian (1988) notamment pour l'estimation d'une matrice variance la plus proche du tableau formé par les covariances calculées pour chaque couple de variables, et pour l'adaptation des formules de régression et de discrimination linéaires lorsque certaines des variables explicatives sont manquantes.

Le deuxième point de vue d'attribution de valeurs aux données manquantes nécessite de se placer dans un cadre paramétrique. Sous cet angle, l'approche la plus répandue et que nous avons adoptée vise à maximiser la vraisemblance de l'échantillon incomplet (cf. Little, Rubin (1987)). Nous reportons sa présentation détaillée au paragraphe 4 où nous présentons également des techniques élémentaires. Avant cela, indiquons une autre approche utilisant l'analyse factorielle (Burtschty, Nora, Vercken 1977). Cette approche, partant de valeurs raisonnables attribuées aux données manquantes, est basée sur un algorithme itératif utilisant alternativement la formule de reconstitution des données à partir des coordonnées factorielles et l'analyse factorielle des données reconstituées.

Enfin, un autre point de vue est souvent adopté, faute de mieux, devant des données manquantes. Il consiste tout simplement à supprimer les individus qui présentent des données manquantes. Dans le cas où le nombre de ces individus est très faible devant le nombre total d'individus, cette pratique est très certainement satisfaisante, mais dans tous les autres cas elle est inappropriée.

4. Les solutions proposées dans SICLA

4.1. Les idées générales

Nous proposons quatre stratégies de complétion de tableaux de données quantitatives ou qualitatives lorsque les données manquantes sont de type DMH. Dans le cas de données quantitatives on fait l'hypothèse qu'il s'agit d'un échantillon d'une variable aléatoire normale multidimensionnelle et dans le cas de données qualitatives on fait l'hypothèse qu'il s'agit d'un échantillon d'une variable aléatoire multinomiale. Nous discuterons au paragraphe 4.4 le bien-fondé de ces hypothèses classiques. Les calculs précis pour l'attribution de valeurs aux données manquantes sont présentés aux paragraphes 4.2 et 4.3 pour ces deux situations. Auparavant, nous énonçons les principes qui leur ont donné naissance.

Les deux premières solutions sont rapides et demandent que de préférence les données manquantes soient de type DMCH.

Solution 1 : on fait l'hypothèse que les variables sont indépendantes, les paramètres des lois marginales de chaque variable sont estimés par le maximum de vraisemblance sur la base des données observées pour la variable concernée ; on attribue alors à chaque donnée manquante la valeur la plus probable. De cette façon, les données manquantes pour une même variable sont affectées d'un valeur unique.

Solution 2 : les hypothèses sont inchangées ; on fait les mêmes calculs pour l'estimation des paramètres de chaque loi marginale, mais on attribue une valeur à chaque donnée manquante par tirage aléatoire suivant la loi marginale de la variable concernée.

Les deux autres solutions ne supposent par l'indépendance des variables.

Solution 3 : les paramètres de la loi multidimensionnelle sont estimés par l'algorithme EM (Dempster, Laird, Rubin (1987) que l'on décrit ici dans son principe. Partant d'une estimation des paramètres de la loi, il s'agit d'un algorithme itératif utilisant alternativement deux étapes. L'étape E (Espérance) consiste à déterminer la loi conditionnelle de chaque donnée manquante sachant les données observées et l'estimation courante des paramètres. L'étape M (Maximisation) consiste à calculer les estimateurs du maximum de vraisemblance des paramètres, les formules faisant usage des lois conditionnelles des données manquantes. De manière naturelle, à la convergence de l'algorithme EM, on attribue à chaque donnée manquante la valeur la plus probable pour l'estimation obtenue des paramètres de la loi multidimensionnelle. De cette façon, tous les individus qui présentent des

données manquantes pour les mêmes variables sont complétés de manière identique.

Solution 4 : les paramètres de la loi multidimensionnelle sont estimés par l'algorithme SEM (Celeux, Diebolt (1987) qui est une version stochastique de l'algorithme EM faisant usage d'un principe d'attribution aléatoire aux données manquantes. Partant d'une estimation des paramètres, il s'agit d'un algorithme itératif utilisant successivement trois étapes. L'étape E est identique à celle de l'algorithme EM. L'étape S (Stochastique) consiste à attribuer une valeur à chaque donnée manquante par tirage aléatoire suivant la loi conditionnelle calculée à l'étape E. L'étape M consiste, alors, à calculer les estimateurs du maximum de vraisemblance des paramètres sur la base de l'échantillon complété de manière aléatoire à l'étape S. Cet algorithme converge en distribution vers une loi de probabilité centrée sur les estimateurs du maximum de vraisemblance, quelle que soit sa position initial. De manière naturelle, à la stationarité de l'algorithme SEM, les valeurs attribuées aux données manquantes sont tirées au hasard suivant les lois conditionnelles de ces données manquantes.

4.2. Les tableaux quantitatifs

Soit N individus décrits par p variables quantitatives. On note $X=(x_i^j;i=1,N;j=1,p)$ le tableau de description des N individus par les p variables. On suppose que le tableau X est incomplet et que les données manquent au hasard. Les stratégies d'attribution de valeurs aux données manquantes sont basées sur l'hypothèse que les N vecteurs de $\mathbf{R}^p$ constituent un échantillon d'une loi normale multidimensionnelle.

Solution 1 : sous l'hypothèse que les p variables sont indépendantes, cette solution conduit à attribuer à chaque donnée manquante la moyenne des valeurs observées pour la variable concernée.

Solution 2 : sous la même hypothèse, la stratégie aléatoire conduit à tirer au hasard chaque valeur attribuée suivant une loi normale réelle dont la moyenne et l'écart-type sont estimés par le maximum de vraisemblance sur la base des valeurs observées pour la variable concernée.

Solution 3 : on suppose maintenant que les p variables sont corrélées. On dispose donc d'un échantillon incomplet d'une loi normale de $\mathbf{R}^p$ de moyenne m et de matrice variance Γ. L'algorithme EM va nous fournir des estimations du maximum de vraisemblance de m et de Γ. dont on déduira la procédure d'affectation déterministe de valeurs aux cases manquantes de X.

Pour décrire l'algorithme EM, il est commode de noter o_i^j les valeurs observées de X et c_i^j les valeurs manquantes, ainsi que de désigner par P_i inclus dans $\{1,..., p\}$ l'ensemble des indices j pour lesquels x_i^j est observé et par Q_j inclus dans $\{1,..., n\}$ l'ensemble des indices j pour lesquels x_i^j est observé. Partant d'une solution (m°, Γ°), une itération de l'algorithme EM qui à (m^n, Γ^n) associe (m^{n+1}, Γ^{n+1}) se décompose ainsi.

Etape E : Calcul des moments conditionnels d'ordre un et deux des valeurs manquantes c_i^j sachant les o_i^j et l'estimation courante (m^n, Γ^n) de (m, Γ). Ces quantités se déduisent des équations classiques donnant les moments conditionnels de vecteurs gaussiens.

$$E\left(c_i^j \mid o, m^n, \Gamma^n\right) = \Gamma_{12}^n(i)\left(\Gamma_{22}^n(i)\right)^{-1}\left(c_i - m_{(i)}^n\right) + m_j^n$$

$$\operatorname{cov}\left(c_i^j, c_i^k \mid o, m^n, \Gamma^n\right) = \Gamma_{11}^n(i) - \Gamma_{12}^n(i)\left(\Gamma_{22}^n(i)\right)^{-1}\Gamma_{21}(i)$$

où

$o = \left(o_i^j;\ j \in P_i;\ i = 1, n\right)$; m_j^n est la $j^{\text{ième}}$ composante de m^n;

$x_i = \left(x_i^{j'}, j' \in P_i\right)$; $m_{(i)}^n = \left(m_{j'}^n, j' \in P_i\right)$;

$\Gamma_{12}(i)$ est le bloc $\left(\Gamma_{k\,l},\ k \notin P_i,\ l \in P_i\right)$ de la matrice Γ;

$\Gamma_{22}(i)$ est le bloc $\left(\Gamma_{k\,l},\ k \in P_i,\ l \in P_i\right)$ de la matrice Γ;

$\Gamma_{21}(i)$ est le bloc $\left(\Gamma_{k\,l},\ k \in P_i,\ l \notin P_i\right)$ de la matrice Γ;

$\Gamma_{11}(i)$ est le bloc $\left(\Gamma_{k\,l},\ k \notin P_i,\ l \notin P_i\right)$ de la matrice Γ;

Etape M : m^{n+1} et Γ^{n+1} sont donnés par les formules

pour $j = 1, p$

$$m_j^{n+1} = \frac{1}{n}\left\{ \sum_{i \in Q_j} o_i^j + \sum_{i \notin Q_j} E\left(c_i^j \mid o, m^n, \Gamma^n\right) \right\}$$

pour $j = 1, p$; $k = 1, p$

$$\Gamma_{jk}^{n+1} = \frac{1}{n}\Big[\sum_{\substack{i \in Q_j \\ i \in Q_k}} \left(o_i^j - m_j^{n+1}\right)\left(o_i^k - m_k^{n+1}\right)$$

$$+ \sum_{\substack{i \in Q_j \\ j \notin Q_k}} \left(o_i^j - m_j^{n+1}\right)\left(E\left(c_i^k \mid o, m^n, \Gamma^n\right) - m_k^{n+1}\right)$$

$$+ \sum_{\substack{i \notin Q_j \\ j \in Q_k}} \left(E\left(c_i^j \mid o, m^n, \Gamma^n\right) - m_j^{n+1}\right)\left(o_i^k - m_k^{n+1}\right)$$

$$+ \sum_{\substack{i \notin Q_j \\ i \notin Q_k}} \Big\{ \left(E\left(c_i^j \mid o, m^n, \Gamma^n\right) - m_j^{n+1}\right)\left(E\left(c_i^k \mid o, m^n, \Gamma^n\right) - m_k^{n+1}\right)$$

$$+ \operatorname{cov}\left(c_i^j, c_i^k \mid o, m^n, \Gamma^n\right) \Big\} \Big]$$

A la convergence de l'algorithme EM on obtient des estimateurs $(\hat{m},\hat{\Gamma})$ de (m,Γ). La stratégie d'attribution de valeurs aux cases manquantes de X consiste alors à affecter tout c_i^j de la valeur $E(c_i^j \mid o, \hat{m}, \hat{\Gamma})$. De manière naturelle, nous initialisons l'algorithme EM par le couple (m°,Γ°) directement déduit de la solution 1 d'attibution de valeurs.

Solution 4 : le modèle est exactement le même que pour la solution 3 mais, ici, l'attribution de valeurs aux cases manquantes se fait par tirage aléatoire. En fait, ces attributions aléatoires peuvent se faire à l'issue de l'algorithme EM. Mais nous avons préféré opter pour l'algorithme SEM dont l'étape S est exactement dans la logique de l'attribution aléatoire. Partant d'une solution initiale (m°,Γ°), une itération de l'algorithme SEM qui à (m^n, Γ^n) associe (m^{n+1}, Γ^{n+1}) se décompose ainsi.

Etape E : calcul des moments conditionnels d'ordre un et deux des valeurs manquantes suivant les mêmes formules que celles énoncées pour l'algorithme EM.

Etape S : toute valeur manquante c_i^j est affectée d'une valeur $(c_i^j)^n$ par tirage aléatoire selon une loi gaussienne de moyenne $E(c_i^j \mid o, m^n, \Gamma^n)$ et de matrice variance $\Gamma_{11}^n(i) - \Gamma_{12}^n(i) \{\Gamma_{22}^n(i)\}^{-1} \Gamma_{21}^n(i)$, ces quantités ayant été calculées à l'étape E. Techniquement, ce tirage aléatoire s'effectue ainsi : on calcule la décomposition de Cholewski TT' de $\Gamma_{11}^n(i) - \Gamma_{12}^n(i) \{\Gamma_{22}^n(i)\}^{-1} \Gamma_{21}^n(i)$, la matrice T étant triangulaire ; on tire au hasard un vecteur u_i à p-card P_i composants selon une loi normale centrée et de matrice variance identité ; la réalisation $(c_i^j)^n$ est alors $T(u_i + E(c_i^j \mid o, m^n, \Gamma^n))$.

Etape M : Cette étape est ici très simple, on estime la moyenne m et la matrice variance Γ par le maximum de vraisemblance sur la base du tableau complété $X^n = ((x_i^j)^n ; i=1,n ; j=1,p)$ où $(x_i^j)^n = o_i^j$ si $i \in Qj$ et $(x_i^j)^n = (c_i^j)^n$ si $i \notin Qj$. Nous ne détaillons pas les formules très classiques qui donnent ainsi naissance aux nouveaux estimés m^{n+1} et Γ^{n+1}.

A la stationnarité, l'algorithme SEM produit donc un tableau complété de manière aléatoire X^n suivant la loi gaussienne estimée. En pratique, on effectue 50 itérations de l'algorithme SEM pour construire le tableau complété.

4.3. Les tableaux qualitatifs

Soit N individus décrits par p variables qualitatives. Chaque variable qualitative j possède m_j modalités que l'on notera $m_1, m_2, \ldots, m_p$. On considère le tableau de données X à N lignes et p colonnes, la case croisant la ième ligne et la jème colonne donnant le numéro de modalité de la variable j pour l'individu i. On suppose que le tableau X est incomplet et que les données manquent au hasard. Les stratégies d'attribution de valeurs aux cases manquantes du tableau X font appel à des hypothèses de distribution multinomiale que nous précisons cas par cas.

Solution 1 : on affecte à chaque valeur manquante la modalité la plus fréquente de la variable concernée. Formellement, cela revient à supposer que les variables sont indépendantes, et qu'elles suivent toutes une loi multinomiale. Dans ce cadre, les probabilités multinomiales de chaque variable sont estimées par les fréquences de ses modalités calculées sur les individus dont les occurences sont disponibles. Ainsi chaque valeur manquante est affectée à la modalité estimée comme la plus probable.

Solution 2 : le modèle est exactement le même, mais cette fois chaque valeur manquante est affectée à l'une des modalités de la variable concernée par tirage au hasard suivant la loi multinomiale dont les paramètres ont été estimés sur la base des individus dont les occurences sont observées.

Solution 3 : on ne suppose plus que les variables sont indépendantes. La procédure d'attribution de valeurs aux données manquantes va tenir compte des interactions entre les variables. Pour la définir précisément nous sommes amenés à considérer l'hypertableau de contingence obtenu par le croisement des p variables et défini ainsi :

$$K = \left\{ k_{j_1,\ldots,j_p};\ j_1=1, m_1;\ldots;j_p=1, m_p \right\}$$

où $k_{j_1,\ldots,j_p}$ est le nombre d'individus présentant simultanément la modalité j_1 de la variable 1,..., la modalité j_p de la variable p.

Il est clair que la connaissance du tableau K définit entièrement le tableau X. On fait l'hypothèse que les N individus constituent un échantillon d'une loi multinomiale d'ordre n et de paramètres

$$p_{j_1,\ldots,j_p};\ j_1=1, m_1;\ldots;j_p=1, m_p$$

$p_{j_1,\ldots,j_p}$ est la probabilité qu'un individu présente simultanément la modalité j_1 de la variable 1,..., la modalité j_p de la variable p. On a bien sûr

$$\sum_{j_1,\cdots j_p} p_{j_1\cdots j_p}=1$$

Pour simplifier les notations, on réindexe de la façon suivante :

$$p_{1,\ldots,1}=p_1;\ p_{1,\ldots,2}=p_2;\ \cdots;\ p_{1\ldots m_p}=p_{m_p};\ \cdots$$

Les paramètres de la loi multinomiale s'écrivent alors (p_ℓ, ℓ=1, M) où M = $\Pi(m_j, j=1, p)$ et la correspondance entre l'indice ℓ et le système d'indiçage initial est donné par la formule

$$\ell = \sum_{k=1}^{p-1} \left\{ (j_k - 1) \prod_{k'=k+1}^{p} m_{k'} \right\} + j_p \qquad \text{(I)}$$

Si les données étaient complètes, les quantités ($k\ell/N, \ell=1,M$) constitueraient les estimateurs du maximum de vraisemblance des ($p\ell$, $\ell=1$, M). Dans le cas d'un tableau incomplet, l'algorithme EM va nous fournir des estimations du maximum de vraisemblance des $p\ell$ dont on déduira une procédure d'attribution déterministe de valeurs aux cases manquantes du tableau X. Soit x le vecteur des valeurs prises sur les p variables qualitatives par un individu quelconque (complet ou incomplet). A x on associe l'ensemble S_x des cases du tableau K à laquelles x peut appartenir. Si x est complet, S_x est bien sûr réduit à une seule case de K. Partant d'une solution initiale $p^\circ = (\overset{\circ}{p}_\ell, \ell=1, M)$, une itération de l'algorithme EM qui à p^n associe p^{n+1} se décompose ainsi.

Etape E : pour tout x et pour toute case $c\ell$ ($\ell=1$, M) de K, on calcule $P(x \in c\ell | p^n)$. Si x est complet

$$P(x \in c\ell | p^n) = \begin{cases} 1 \text{ si } c\ell \in S_x \\ 0 \text{ sinon} \end{cases}$$

si x est incomplet

$$P(x \in c\ell | p^n) = \begin{cases} p_\ell^n \Big/ \sum_{c\ell' \in S_x} p_{\ell'}^n \text{ si } c\ell \in S_x \\ 0 \text{ sinon} \end{cases}$$

Etape M : $p_\ell^{n+1} = \sum_x P(x \in c\ell | p^n)$ pour $\ell = 1, M$.

A la convergence de l'algorithme EM on obtient des estimateurs $\hat{p} = (\hat{p}\ell, \ell=1, M)$ des ($p\ell$, $\ell=1$, M).

Maintenant la stratégie d'attribution des valeurs les plus probables aux cases manquantes du tableau de données X conduit à affecter tout vecteur incomplet x à la case $c\ell$ du tableau K telle que

$$k\ell = \sup_{\ell'} (k\ell' \text{ avec } c\ell' \in S_x)$$

On en déduit aisément le vecteur de description complété $\hat{x} = (j_1,\ldots,j_p)$ d'après la formule (I) par l'algorithme de la division euclidienne :

- j_p est le reste de la division de ℓ par m_p si ce reste est non nul, j_p est m_p sinon.
- j_{p-1} est le reste, augmenté de un, de la division de $(\ell-j_p)/m_p$ par m_{p-1}.
- etc.

De manière naturelle, nous initialisons l'algorithme EM par les p°_{ℓ} déduits de la solution 1 d'attribution de valeurs.

Solution 4 : le modèle est exactement le même que pour la solution 3 mais, ici, l'attribution de valeurs aux cases manquantes se fait par tirage aléatoire. Là aussi, ces attributions aléatoires peuvent se faire à partir de $\hat{p}_{\ell}$ obtenues à l'issue de l'algorithme EM, mais nous avons pour les même raison qu'en 4.2 opté pour l'algorithme SEM. Partant d'une solution initiale $p^{\circ} = (p^{\circ}_{\ell}, \ell=1, M)$ une itération de l'algorithme SEM qui à p^n associe p^{n+1} se décompose ainsi.

Etape E : cete étape est la même que celle de l'algorithme EM.

Etape S : tout vecteur incomplet x est attribué à l'une des cellules c_{ℓ}, $c_{\ell} \in S_x$, par tirage aléatoire suivant la loi multinomiale de paramètres

$$\left(\frac{p^n_{\ell}}{\sum_{c_{\ell'} \in S_x} p^n_{\ell'}}, c_{\ell} \in S_x \right)$$

Etape M : a l'issue de l'étape S, le tableau K est complété en un tableau $K^n = (k^n_{\ell}, \ell=1,M)$ et on obtient $p^{n+1}_{\ell} = k^n_{\ell} / N$ pour $\ell=1,M$.

A la stationnarité, l'algorithme SEM produit donc un tableau de données complétée X^n déduit directement du tableau K^n. En pratique on effectue 50 itérations de l'algorithme SEM pour construire le tableau complété.

4.4. Commentaires

Nous examinons ici les caractéristiques et les limites des stratégies proposées.

Les solutions 1 et 2 sont valides en toute rigueur si l'hypothèse d'indépendance des variables est acceptable, ce qui est rarement le cas. Cependant, d'un point de vue pratique, elles peuvent être vues comme des techniques rapides et raisonnables d'attribution de valeurs aux données manquantes, en particulier lorsqu'elles sont de type DMCH. De fait la solution 1 est la technique la plus couramment proposée dans les logiciels statistiques.

Les solutions 1 et 3 attribuent des valeurs "centrales" aux données manquantes et conduisent ainsi à sous-estimer les variances de l'échantillon complété. Ce biais dans l'estimation des variances risque de ne pas être négligeable si le nombre de données manquantes est important. L'un des principaux intérêts des solutions 2 et 4 est précisément de ne pas sous-estimer les variances de l'échantillon complété.

L'algorithme EM est un algorithme très utilisé dans de nombreux domaines pour trouver les estimateurs du maximum de vraisemblance à partir de données incomplètes. Néanmoins, il peut présenter des défauts importants : ses résultats peuvent dépendre fortement de sa position initiale ; il peut converger extrémement lentement ; il peut converger vers un col de la vraisemblance. Dans le cadre où nous l'utilisons, ces défauts ne sont pas réellement présents : le choix de sa position initiale s'effectue de manière naturelle ; il ne rencontre pas de situation de convergence lente pour peu que le nombre de données manquantes reste raisonnable. Dans les exemples sur des données simulées que nous avons traités, au même titre que l'algorithme SEM, il fournissait des estimations précises des paramètres de la loi considérée. On aurait pu sans inconvénient utiliser le principe d'attribution aléatoire de valeurs pour compléter le tableau à l'issue de l'algorithme EM. Il nous est apparu plus logique d'effectuer cette attribution aléatoire à l'issue de l'algorithme SEM. De plus l'algorithme SEM a l'avantage de ne pas renconter de situation de convergence lente contrairement à l'algorithme EM (même si, ici, ce genre d'incident est très rare).

En résumé, les algorithmes EM et SEM apparaissent quasi équivalents pour estimer les paramètres de la loi, mais la solution 4 fournit à partir du tableau complété aléatoirement des estimations des variances non sous-estimées.

Maintenant, on peut s'interroger sur le bien-fondé de l'hypothèse de normalité pour les tableaux quantitatifs ou sur l'hypothèse de loi multinomiale pour les tableaux qualitatifs. Rappelons tout d'abord la nécessité de faire des hypothèses précises sur la loi de l'échantillon incomplet pour définir une procédure d'estimation cohérente des données

manquantes. D'autre part, les tableaux complétés sont destinés à être l'objet d'analyses des données multidimensionnelles. Pour les données quantitatives, la plupart des méthodes relèvent de l'analyse linéaire (régression, discrimination, analyse en composantes principales) ou utilisent des critères d'inertie pour la classification. Ainsi ces techniques s'appuient sur, ou du moins ne remettent pas en cause, l'hypothèse de normalité des données. Pour les données qualitatives, de nombreuses méthodes disponibles sous SICLA font implicitement l'hypothèse que les données suivent une loi multinomiale (analyse des correspondances, classification avec le critère du χ^2, discrimination). Pour les techniques de classification, on trouvera un exposé des liens entre critère d'inertie et normalité, d'une part, et critère du χ^2 et les lois multinomiales, d'autre part, dans Celeux (1988). En résumé, les hypothèses de travail faites pour compléter un tableau de données sont cohérentes vis-à-vis des méthodes d'analyse des données de SICLA.

Il n'existe pas actuellement dans SICLA de techniques de complétion qui tiennent compte simultanément de données manquantes pour des variables quantitatives et pour des variables qualitatives. On trouvera dans Little, Rubin (1987), chapitre 10, des techniques pour traiter les données manquantes dans un tel cas.

5. Les commandes DOMAQN et DOMAQL

La commande DOMAQN crée une structure des données complétée à partir d'une structure de données comportant des valeurs manquantes pour les variables quantitatives de type mesure. La complétion des données se fait suivant l'une des quatre solutions présentées au paragraphe 4.2. De plus, on édite les moyennes et les variances des variables pour la structure de données initiale. Ces caractéristiques sont calculées en remplaçant les valeurs manquantes selon la solution 1. Pour les solutions 2, 3 et 4, on édite également les moyennes et les variances des variables pour la structure de données complétée.

La commande DOMAQL crée une structure de données complétée à partir d'une structure de données comportant des valeurs manquantes pour les variables qualitatives. La complétion des données se fait suivant l'une des quatre solutions présentées au paragraphe 4.3. De plus, on édite les répartitions des modalités des variables de la structure initiale, et, à la suite, on édite ces répartitions pour la structure de données complétée.

Enfin, signalons que les commandes de SICLA de description élémentaire peuvent fonctionner sur une structure de données comportant des données manquantes (non complétée). Pour ce faire, les commandes DESQAN, DESQAL, BOX, PLOT, AKHI2 suppriment les individus

comportant des données manquantes, et les commandes HISTO et CORREL utilisent la solution 1 du paragraphe 4.2.

Références

Carbon M., Gourieroux C., Huber C., Lecoutre J.P. (1989) "Analyse des données de durée de vie" ed. Economica (à paraitre).

Celeux G., (1988) "Classification et modèles" R.S.A. Vol. 36 n° 3.

Celeux G., Diebolt J. (1987) "A probabilistic teacher algorithm for iterative maximum likelihood estimation" IFCS 87. North-Holland p. 617-623.

Dempster A.P., Laird N.M., Rubin D.B. (1977) "Maximum likelihood from incomplete data via the EM algorithm" J.R.S.S.B. 39 p 1-38.

Der Megreditchian G. (1988) "Problèmes engendrés par les données manquantes dans la pratique statistique" Rapport EERM N° 208.

van der Heijden P.G.M., Escofier B. (1988) "Multiple corresponding analysis with missing data" Rapport de recherche INRIA N° 902.

Little R.J.A., Rubin D.B. (1987) "Statistical analysis with missing data" Wiley.

Kalbfleisch J.D., Prentice R.L. (1980) "The statistical analysis of failure time data" Wiley.

Lebart L., Morineau A., Tabard N. (1977) "Techniques de la description statistique" Dunod.

Rubin D.B. (1986) "Multiple Imputation for non response in surveys" Wiley.

SICLA (1988) " Manuel de référence" INRIA.

Imprimé en France
par
l'Institut National de Recherche en Informatique et en Automatique