Chapitre 4 : Problème de données manquantes dans un tableau numérique. Une application de l'A.S.I.

Régis Gras

Equipe COnnaissances & Décision (COD)
Laboratoire d'Informatique de Nantes Atlantique – FRE CNRS 2729
Site Ecole Polytechnique de l'Université de Nantes
La Chantrerie BP 50609 44306 Nantes cedex 3
regisgra@club-internet.fr

Résumé. Une base de données croisant variables et sujets issues d'observations présente souvent des vides dus, par exemple, à des absences de réponse ou à l'impossibilité matérielle de la recueillir. Or, pour en effectuer un traitement, il est essentiel de disposer d'un tableau complet. L'analyse statistique implicative, entre autres méthodes d'analyse de données, à l'œuvre au moyen du logiciel de traitement C.H.I.C. (Couturier et Gras., 2005), impose que les vides soient comblés. Se pose alors le problème de déterminer quelle valeur la plus vraisemblable attribuer à la variable non observée sur tel sujet ou, de façon symétrique, quelle valeur attribuer à un sujet sur une variable donnée et muette sur lui. Nous présentons ici une méthode qui, au vu du comportement de réponse observé par le sujet sur d'autres variables, intimement liées à la variable muette, permet de pallier la carence locale. Un exemple numérique illustre l'usage de cette méthode sur un tableau incomplet.

1 Problématique et contraintes sémantiques

L'analyse statistique implicative traite un tableau de données qui croise un ensemble de variables V, colonnes du tableau, de cardinal v et un ensemble E de sujets (ou d'objets), lignes du tableau, de cardinal n. L'intersection d'une ligne et d'une colonne représente donc la valeur x(i) prise par l'individu x selon la variable i. Dans le cas où les variables V sont binaires x(i) = 0 ou 1. Dans les autres cas classiques (variables modales, fréquentielles, numériques normées), les valeurs sont des nombres réels de l'intervalle [0,1].

Or il est possible que les observations ou les mesures faites conduisant au tableau présentent des "trous", c'est-à-dire des absences de réponse. Il existe donc des individus x et des variables i telles que la **valeur x(i) soit manquante**. Si l'on ne souhaite pas supprimer l'individu x, donc toutes les autres observations faites en x selon les autres variables, et par suite perdre un certain nombre d'informations permettant d'étudier les relations implicatives entre variables, le problème va consister à choisir la valeur la plus pertinente que l'on pourra affecter à x(i) pour la structure des variables sous-jacente.

L'idée intuitive est de chercher quels sont les individus qui ont des comportements semblables à x selon les autres variables sur lesquelles nous disposons d'informations pour x et d'attribuer à x(i) la valeur correspondant à celle que prend en i l'individu y, le plus "ressemblant" à x. La modélisation va donc porter sur le choix d'un critère de ressemblance entre individus.

De nombreuses méthodes[1] visant à remplacer des données manquantes par des valeurs estimées existent. Le logiciel libre R permet d'y parvenir en utilisant des procédures de régression ou de recherche du maximum de vraisemblance. Nous présentons ici deux méthodes originales : l'une exploite la proximité relative des individus eu égard à leur instanciation des variables ; l'autre exploite les relations implicatives entre les variables à données complètes.

2 Contraintes analytiques sur le modèle

Nous supposons que x n'est pas défini en i alors que y l'est par la valeur y(i). Nous supposons également que le nombre de valeurs non définies en x n'est pas trop grand et que x et y prennent des valeurs sur de nombreuses variables communes, faute de quoi l'estimation de x(i) serait entachée d'imprécision. Cette hypothèse est assez générale pour ce qui suit, le risque d'erreur croissant avec le nombre de « trous » dans le tableau de données.

Le critère de ressemblance entre deux individus x et y doit être nécessairement symétrique comme le sont tous les critères de similarité.

Il doit prendre en compte le maximum d'informations sur les valeurs prises par x et y selon les variables qu'ils affectent en commun. Ce sont donc les profils de x et y sur ces variables qu'il nous faudra envisager. Ces profils n'intégreront donc que les variables où, pour x **et** y, il n'y a pas de valeurs manquantes.

Ces profils doivent prendre en compte toutes les valeurs en jeu en intégrant la qualité de représentation relative des individus selon l'ensemble des variables sur lequel ils sont observés. Si, par exemple, la valeur x(j) est élevée et si la somme totale marginale $x. = \sum_{j \in V} x(j)$ des valeurs prises par x sur V (ou une partie de V en cas de valeurs manquantes pour x) est relativement faible, on accentuera la position de j dans le profil en calculant $\frac{x(j)}{x.}$

3 Méthodologie de substitution

Le profil de x sera donc l'ensemble $\left\{ \frac{x(j)}{x.} \right\}_{j \in V}$, distribution conditionnelle de j sachant x. L'écart entre les termes $\frac{x(j)}{x.}$ et $\frac{y(j)}{y.}$ du profil observé en x et y selon une variable j quelconque doit également être relativisé à la représentation de cette variable sur l'ensemble des individus. Si, par exemple, une variable ne prend que de faibles valeurs sur E, un écart important entre $\frac{x(j)}{x.}$ et $\frac{y(j)}{y.}$ doit être majoré par rapport à ce qu'il serait si la variable

[1] Voir par exemple : Schafer J.L. (2000), *Analysis of incomplete multivariate data*, Chapman and Hill.

n'avait pris que de fortes valeurs. On satisfera cette contrainte en divisant chaque distorsion entre les termes des profils par la somme marginale $n_j = \sum_{z \in E} z(j)$. La division de n_j par N ne change nullement le rapport entre les distorsions mais permet d'homogénéiser ce rapport en des termes relatifs. Un tableau initial se présente donc comme ceci :

V→ E↓	a	…	i	…	j	…	Marge
1	1(a)	…	1(i)	…	1(j)	…	1.
…	…	…	…	…	…	…	…
x	x(a)	…	XXXXX	…	x(j)	…	x.
…	…	…	…	…	…	…	…
y	y(a)	…	y(i)	…	y(j)	…	y.
…	…	…	…	…	…	…	…
Marge	n_a	…	n_i	…	n_j	…	N

TAB. 1

La dissemblance totale prend alors la forme d'une distance du χ^2 entre les deux profils relatifs de x et y, comparable à celle utilisée en analyse factorielle des correspondances, soit :

$$d(x,y) = \left[\sum_{j \in V', j \neq i} \frac{\left[\frac{x(j)}{x.} - \frac{y(j)}{y.} \right]^2}{\frac{n_j}{N}} \right]^{\frac{1}{2}}$$

où V' est le sous-ensemble de V des variables selon lesquelles x et y prennent en commun des valeurs. V-V' est donc l'ensemble des variables où x et/ou y présentent l'un et/ou l'autre des valeurs manquantes. Autrement dit, le calcul se fait sur le tableau extrait du tableau initial duquel ont été supprimées les colonnes relatives aux variables de V-V', y compris le calcul des valeurs marginales (individus et variables).

Cependant, un autre facteur est "aggravant" par rapport à cette dissemblance. En effet pour que l'ensemble des dissemblances entre x et les autres individus permette des comparaisons, il faut prendre en compte le cardinal k de V'. Si ce cardinal est faible, la dissemblance définie ci-dessus risque d'être faible et peu informative. On l'affecte donc d'un coefficient correcteur qui décroît si le nombre k de variables partagées en commun par x et y est important, soit le rapport : $\frac{v-k}{v}$. Par conséquent , on pose :

$$\delta(x,y) = \frac{v-k}{v} d(x,y)$$

En définitive, on affectera à x(i) la valeur y_0(i) si $\delta(x,y_0) = \inf_{y \in E / \{x\}} \delta(x,y)$. Si deux individus satisfont ce minimum, on choisira, prioritairement, celui qui est relatif à un sous-

ensemble V' le plus important. En cas de nouvelle égalité, on choisira celui dont l'effectif marginal y. est le plus fort. En cas de nouvelle égalité, on retiendra arbitrairement l'individu le plus haut placé dans le tableau ExV d'entrée.

A la suite de chaque affectation de valeurs estimées, les marges sont remises à jour, de même que la valeur de N.

Remarque 1

Il est donc évident que l'ordre dans lequel sont effectuées les opérations de substitution aura une influence sur les estimations successives. Leur rareté supposée nous permet de faire l'hypothèse que cette influence sera négligeable.

Remarque 2

On rencontre dans des situations d'évaluation de travaux d'élèves des épreuves à plusieurs modalités. Par exemple, pour des raisons de sécurité vis-à-vis du copiage, mais également pour des mises à l'épreuve d'hypothèses sur la proximité de complexité d'items, on partitionne une population en 3 groupes X, Y et Z. Chacun de ces groupes est soumis à deux cahiers d'items parmi trois cahiers A, B et C comme l'indique le tableau suivant :

X	A	B	XXXX
Y	XXXX	B	C
Z		XXXX	C

TAB.2

L'absence de résultats des élèves de X en C sera compensée par une estimation à partir de la comparaison de ses comportements de réponse selon B avec les élèves de Y et selon A avec les élèves de Z. On prendra, par exemple, la moyenne des estimations faites à partir de Y et de Z. S'il s'agissait de variable binaire (réussite-échec), on choisirait l'entier 0 ou 1 le plus proche de cette moyenne.

Remarque 3

On aurait pu utiliser le coefficient de corrélation linéaire pour signifier la ressemblance entre les individus x et y. Cependant, nous lui préférons l'indice défini plus haut qui intègre, mieux que la corrélation, les nuances distributionnelles prises en compte dans sa définition.

De même, on aurait pu faire appel au coefficient de similarité de I.C. Lerman (1981). Mais celui-ci ne paraissait pas permettre la relativisation que nous avons introduite.

4 Une autre approche dans le cadre de l'A.S.I.

Un autre point de vue mettant l'accent sur les variables plutôt que sur les individus est également envisageable dans le cadre de l'Analyse Statistique Implicative (Gras et al. 2001 et Gras, 2005) et Partie 1, chap 1. Ainsi, l'absence d'une donnée a(x) selon la variable a peut être compensée par l'attribution d'une valeur statistique en considérant la "proximité implicative" de a avec des variables dont on connaît la valeur prise en x. Par exemple, si C_a et C'_a sont respectivement les classes de variables impliquées par a et impliquant a au seuil $1-\alpha$, il suffit de considérer leur intersection avec les classes associées respectivement à celles des variables b pour lesquelles b(x) est connu. Un indice de distance, comparable à celui que nous avons défini plus haut, permettrait d'affecter un a(x) une valeur la plus vraisemblable eu

égard aux valeurs de type b(x). Le logiciel d'analyse de données CHIC (Classification Hiérarchique Implicative et Cohésitive) (Couturier et Gras. 2005) et Partie 2, chap.11 permet de mettre en évidence aisément les variables associées à a (resp. b) en choisissant l'option « Sélectionne les variables en mode cône ». La variable a sélectionnée présente, seule, ses pères et ses fils. La métaphore « cône » est justifiée par la structure qui apparaît comme le ferait un cône à deux nappes. Si, à son tour, la variable b est sélectionnée, on peut, avec un seuil convenable, examiner l'intersection des deux cônes.

Il est possible d'opérer autrement et d'utiliser les liaisons implicatives entre les variables d'une autre façon qui permette d'enrichir l'estimation de la donnée manquante x(i), faite précédemment.

Pour cela, considérons le schéma implicatif établi autour de la variable i et n'intégrant pas l'action de la donnée manquante en x mais celles des mêmes sujets. Dans le cas général, pour un seuil de l'intensité d'implication fixé (par exemple 0.95), une certaine classe de variables C, classe père disposant de valeurs en x, non totalement ordonnées, établit des liaisons implicatives vers i ; de même, une autre classe C', classe fils disposant également de valeurs en x, non totalement ordonnées, est constituée de variables impliquées par i. Si l'une des classes C et C' est vide, ce qui suit n'intéressera que celle qui est non vide. Si les deux sont vides et si un abaissement du seuil de l'intensité fait perdre toute signification , on conserve les résultats de l'estimation de x(i) précédente.

Revenons au cas général et retenons alors les variables j et k qui présentent des intensités d'implication maximales, respectivement de j de la classe C vers i et de i vers k de la classe C'. Ainsi, la variable j implique de façon la plus consistante au sein de C la variable i qui elle-même implique de façon la plus consistante au sein de C' la variable k. C'est à ces variables que l'on accordera le meilleur crédit pour affiner l'estimation de x(i).

L'une d'entre ces deux classes présente une cohésion plus forte ou égale à celle de l'autre. S'il s'agit de la classe C, nous choisirons pour x(i) le minimum de la valeur précédemment établie et de la valeur x(j) observée en x selon j, minimum qui permet de ne pas détruire, le plus souvent, la liaison de j vers i. S'il s'agit de la classe C', nous choisirons pour x(i) le maximum de la valeur précédemment établie et de la valeur x(k) observée en x selon k, maximum qui permet également de ne pas détruire, le plus souvent, la liaison de k vers i.

Ainsi, nous disposons d'une information plus complète pour déterminer la valeur manquante puisque nous intégrons simultanément et les comportements de x comparés à ceux des autres sujets relativement aux variables et ceux des variables dans l'ensemble de leurs liaisons au sein de la population où elles sont comparables.

5 Exemple didactique

Supposons donné le tableau suivant correspondant à l'observation sur 4 individus de 4 variables modales a, b, c et d , prenant les valeurs 0 ou 0,25 ou 0,5 ou 0,75 ou 1.

Il s'agit donc d'estimer les valeurs e3(c) et e4(d) sur la base des autres observations. Les marges de ce tableau n'intègrent que les variables où, dans un premier temps, e3 est estimé à l'aide de e1, e2 et e4. Donc, pour cette estimation, la variable c n'intervient pas pour e1, e2 et e4.

V → / E ↓	a	b	c	d
e1	0	0,50	0,25	0
e2	0,50	0,75	1	0,25
e3	0,50	0,50	**?**	0,25
e4	0,75	0,50	1	?

TAB.3 *: Tableau d'entrée des données*

V→ / E ↓	a	b	c	d	Marge pour e3
e1	0	0,50	0,25	0	0,50
e2	0,50	0,75	1	0,25	1,50
e3	0,50	0,50	**?**	0,25	1,25
Marges	1	1,75	1	0,50	3,25

TAB. 4 : *Tableau pour le calcul des distances de e3 à e1 et e2*

V→ / E ↓	a	b
e1	0	1
e2	1/3	1/2
e3	0,40	0,40

TAB. 5 *Fréquences conditionnelles des individus*

Les calculs successifs donnent les résultats suivants pour v=4 :

d(e1, e3) = 1,203 et δ (e1, e3) = 0,301 (pour k=3)

d(e2, e3) = 0,237 et δ (e2, e3) = 0,0592 (pour k=3)

TAB.6

Par exemple :

$$d(e1,e3) = \sqrt{\frac{(0-0,4)^2}{\frac{1}{3,25}} + \frac{(1-0,4)^2}{\frac{1,75}{3,25}} + \frac{(0-0,2)^2}{\frac{0,5}{3,25}}} = 1,203 \quad \text{d'où} \quad \delta(e1,e3) = \frac{1,203}{4} = 0,301$$

car k=3. Dans la comparaison de e3 avec e4, les variables c et d n'interviennent pas (variables non communes).

V→ / E↓	a	b	Marges
e3	0,50	0,50	1
e4	0,75	0,50	1,25
Marges	1,25	1	2,25

TAB. 7 *Tableau pour le calcul de distance entre e3 et e4*

V→ / E↓	a	b
e3	0,50	0,50
e4	0,60	0,40

TAB. 8 : *Fréquences conditionnelles de e3 et e4*

d'où d(e3, e4) = 0,201 et δ (e3, e4) = 0,101 (pour k= 2) alors que δ (e1, e3)= 0,301 et que δ (e2, e3) = 0,0592.

Par suite, e2 est l'individu le plus "ressemblant" à e3. On choisira donc l'estimation :

e3(c) = e2(c) = 1.

On obtient alors une nouvelle distribution donnée par les tableaux suivants à partir desquels nous estimerons e4(d) :

V→ E↓	a	b	c	d	Marges pour e4
e1	0	0,50	0,25	(0)	0,75
e2	0,50	0,75	1	(0,25)	2,25
e3	0,50	0,50	**1**	(0,25)	2
e4	0,75	0,50	1	?	2,25
Marge	1,75	2,25	3,25	?	

TAB. 9 *Tableau de données après estimation de e3(c)*

V→ E↓	a	b	c
e1	0	2/3	1/3
e2	2/9	3/9	4/9
e3	0,25	0,25	0,5
e4	3/9	2/9	4/9

TAB.10 : *Fréquences conditionnelles des individus*

Cette fois, toutes les informations distributionnelles relatives à d disparaissent puisque e4 n'est pas défini en d. Reprenons les calculs en tenant compte de la première estimation de e3(c). Nous obtenons :

d(e1, e4) = 1,060 et δ (e1, e4) = 0,265 (pour k = 3)

d(e2, e4) = 0,301 et δ (e2, e4) = 0,075 (pour k = 3)

d(e3, e4) = 0,397 et δ (e3, e4) = 0,099 (pour k= 3)

Par suite, e2 est l'individu le plus "ressemblant" à e4. On choisira donc l'estimation : e4(d) = e2(d) = 0,25. On obtient alors un tableau complet sur lequel on pourra pratiquer l'analyse statistique implicative habituelle.

V→ E↓	a	b	c	d
e1	0	0,50	0,25	0
e2	0,50	0,75	1	0,25
e3	0,50	0,50	**1**	0,25
e4	0,75	0,50	1	**0,25**

TAB. 11

6 Conclusion

Nous avons présenté une méthode numérique afin de pallier les carences d'un tableau de données incomplet. A l'aide d'une distance de type χ^2 intégrant le maximum d'informations observées sur les variables sur lesquelles des sujets sont mesurés, nous avons affecté à un sujet, incomplètement défini selon une variable, une valeur qu'il aurait pu prendre sur celle-ci, eu égard à sa ressemblance avec les autres sujets selon les variables partagées en commun. L'exemple numérique, traité à la main, illustre cette méthode. Nous avons aussi évoqué une autre méthode qui s'appuierait sur les données fournies par les calculs des implications respectives entre les variables partout où elles sont complètement définies. Cette stratégie qui prend plus encore en compte la structure implicative de l'ensemble des variables sera développée par ailleurs.

Références

Agrawal R., Imielinski T. and Swami A. (1993) : Mining association rules between sets of items in large databases. In Peter Buneman and Sushil Jalodia, Editors, *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data, p.207-216, Washington, D.C. 26-28*

Couturier R, Gras R. (2005) : CHIC : Traitement de données avec l'analyse implicative, *Extraction et Gestion des Connaissances, Volume I1, RNTI, Cepadues, Paris, p.679-684, , ISBN 2.85428.683.9*

Gras R., Kuntz P. et Briand H.(2001): Les fondements de l'analyse statistique implicative et quelques prolongements pour la fouille de données, *Mathématiques et Sciences Humaines*, n° 154-155, p 9-29, ISSN 0987 6936

Gras R. (2005): Panorama du développement de l'A.S.I. à travers des situations fondatrices, *Actes de la 3ème Rencontre Internationale A.S.I., Supplément n° 15 de la Revue « Quaderni di Ricerca in Didattica* », *p. 9-33, ISSN 1592-5137*, Université de Palerme

Summary

The problem of missing data is a well known problem in data analysis. The Statistical Implicative Analysis (implemented in the software C.H.I.C. (Couturier and Gras, 2005)), like others data mining methods, needs that all the variables of the dataset are instantiated. Therefore, the problem is to set the missing variables with the most likely values. We present in this chapter a method to handle this problem, using information about the values of others variables, correlated with the missing one, on the same example. A numerical example is given to illustrate our method.