REVUE DE STATISTIQUE APPLIQUÉE

F. CRETTAZ DE ROTEN
J.-M. HELBLING

Une estimation de données manquantes basée sur le coefficient RV

Revue de statistique appliquée, tome 39, n° 2 (1991), p. 47-57
<http://www.numdam.org/item?id=RSA_1991__39_2_47_0>

Rev. Statistique Appliquée, 1991, **XXXIX** (2), 47-57

UNE ESTIMATION DE DONNÉES MANQUANTES BASÉE SUR LE COEFFICIENT RV.

F. CRETTAZ DE ROTEN, J.-M. HELBLING

Département de mathématiques,
Ecole Polytechnique Fédérale de Lausanne (Suisse)

RÉSUMÉ

Dans cet article, on propose une nouvelle méthode d'estimation des données manquantes de type analyse des données. Pour l'estimation, on utilise la structure multivariée des données en maximisant le coefficient RV (Escoufier 1973) exprimé en fonction du vecteur manquant. La méthode développée dans la première partie, sera confrontée aux méthodes existantes sur la base d'un exemple. Les deux critères définis par Gleason et Staelin en 1975, permettront d'évaluer les performances de chacune des méthodes.

Mots-clés : *données manquantes en analyse multivariée, estimation des données manquantes, coefficient RV*

1. Introduction

En inférence statistique multivariée et en analyse de données multivariées, on rencontre fréquemment des données manquantes : refus de répondre à certains items d'un questionnaire, appareil de mesure qui tombe en panne, etc. Elles posent un dilemme au statisticien car, soit il résout de manière abrupte le problème en éliminant le ou les individus ayant des données manquantes pour une ou plusieurs variables, soit il se penche sur l'abondante littérature sur les données manquantes qui, en lui proposant plusieurs techniques, l'obligera à faire certains choix délicats.

En premier lieu, le statisticien devra préciser la nature de ses données manquantes. Pour un individu, une variable peut être manquante indépendamment de sa valeur et de la valeur d'autres variables, on parle alors de Donnée Manquante Complètement au Hasard (notée DMCH). Lorsque la perte de la donnée est indépendante de sa valeur mais dépendante de la valeur d'autres variables, on parle de Donnée Manquante au Hasard (noté DMH). Dans les autres cas, on parle de donnée manquante non au hasard. Cette classification, définie par Rubin en 1976, est admise actuellement. Il est très délicat de traiter des données qui ne sont pas de type DMCH, c'est pourquoi on fait souvent cette hypothèse. La vérification de cette hypothèse est soit empirique (connaissance des données et du processus de récolte), soit assez complexe (Simon et Simonoff 1986, Little 1988).

En second lieu, il devra choisir entre deux approches :

1) L'*estimation de paramètres* avec données manquantes. De nombreuses méthodes multivariées s'appuyent sur une réduction initiale des données au vecteur des moyennes et à la matrice de variance-covariance, c'est pourquoi cette approche consiste à estimer ces quantités à partir des données incomplètes (Timm 1970, Dempster, Laird et Rubin 1977, Curry et Kim 1977, Kariya, Krishnaiah et Rao 1983, Der Megreditchian 1988). Il n'est donc, dans cette approche, pas réellement nécessaire d'estimer les données manquantes elles-mêmes, puisque moyennes et variances-covariances estimées permettent d'effectuer les techniques statistiques classiques.

2) L'*estimation directe* des données manquantes que la littérature anglaise appelle imputation des données manquantes. Cette approche complète les données avant toute utilisation, en estimant la valeur des données manquantes. Les principales méthodes de ce type se basent :

- sur les moyennes : la valeur manquante est remplacée par un estimateur de l'espérance de la variable concernée,
- sur la régression simple ou multiple : on estime la donnée manquante par la valeur que prend la variable concernée lorsqu'on la régresse sur une ou plusieurs variables (Buck 1960, Frane 1976, Seber 1984),
- sur la modélisation des données à l'aide de la loi normale et sur la fonction de vraisemblance (Srivastava 1985, Little et Rubin 1987),
- sur l'analyse en composantes principales (Dear 1959).

Les techniques basées sur la régression, souvent appelées *méthodes de Buck*, comportent de nombreuses variantes liées aux diverses façons de calculer les paramètres de la régression.

L'algorithme itératif EM se situe à la frontière des deux approches puisqu'alternativement l'étape E estime les données manquantes et l'étape M estime par maximum de vraisemblance les paramètres (Dempster, Laird et Rubin 1977, Boyles 1983, Little et Rubin 1987).

Dans la littérature, certains articles comparent un sous-ensemble des méthodes précédentes au niveau théorique (Frane 1976, Little et Rubin 1987), d'autres optent pour une comparaison expérimentale, soit en simulant des données normales (Curry et Kim 1977, Kaiser 1990) et parfois en plus en se plaçant dans un contexte de régression (Haïtovsky 1968, Little 1979), soit en utilisant des données réelles (Buck 1960). Enfin, certains auteurs réunissent la comparaison théorique et empirique (Gleason et Staelin 1975, Beale et Little 1975).

Récemment, une nouvelle conception du traitement des données manquantes a vu le jour, elle consiste à remplaçer une donnée manquante par un ensemble de M valeurs ($M > 1$) créant ainsi M tableaux complétés. L'estimation multiple (imputation multiple en anglais) utilise les méthodes standard pour fournir les M estimations et pour analyser chaque tableau complété puis, calcule l'effet des diverses estimations des données manquantes (Herzog et Rubin 1983, Rubin et Schenker 1986, Rubin 1987).

Finalement, d'autres méthodes sont liées à une situation particulière : analyse de variance (Anderson 1946, Rubin 1972, Dodge 1985, Murray 1986), enquête (Ford 1983, Madow, Olkin et Rubin 1983, Grosbras 1987, Little 1986, Rubin 1987), séries temporelles (Jones 1980, Abraham 1981),... Le livre de Little et Rubin, paru en 1987, traite brièvement ces situations et fournit des références supplémentaires.

L'analyse des méthodes existantes soulève des questions. Peut-on faire des hypothèses si lourdes (de modèle ou de distribution)? Ne devrait-on pas rester dans un contexte d'analyse des données? Les désavantages connus de ces méthodes (sous-estimation de la variabilité, distorsion de la distribution) sont-ils acceptables?

Dans cet article, nous allons présenter une nouvelle technique d'estimation des données manquantes : elle utilise la structure multivariée des données et se base sur la maximisation du coefficient RV introduit par Escoufier en 1973. La définition et les propriétés du RV ainsi que de la nouvelle méthode d'estimation seront étudiés au paragraphe 2. Dans le paragraphe 3, nous allons comparer sur la base d'un exemple les principales méthodes d'imputation avec la méthode RV à l'aide de deux critères définis par Gleason et Staelin en 1975.

2. Une nouvelle méthode d'estimation de données manquantes

2.1 L'idée de la méthode

La méthode la plus utilisée est probablement la méthode de Buck qui attribue à la valeur manquante la prédiction fournie par la régression de cette variable sur les autres variables. On peut montrer qu'elle revient à estimer la donnée manquante par la valeur qui minimise la distance de Mahalanobis entre le vecteur individu qui contient la donnée manquante et le vecteur moyenne. Cette minimisation d'une fonction d'une norme vectorielle nous a donné l'idée de minimiser une fonction d'une norme matricielle pour se placer dans un contexte multivarié général.

2.2 Le coefficient RV

Soit $X = \begin{pmatrix} X^{(1)} \\ X^{(2)} \end{pmatrix}$ un vecteur aléatoire tel que $X^{(1)}$ est un vecteur $p \times 1$ et $X^{(2)}$ est $q \times 1$. La séparation en deux parties du vecteur X se fait de façon naturelle par exemple, variables endogènes et exogènes, ou qualités physiques et intellectuelles. Sa moyenne μ et sa matrice de variance-covariance Σ sont :

$$\mu = E(X) = \begin{pmatrix} \mu^{(1)} \\ \mu^{(2)} \end{pmatrix} \quad \text{et} \quad \Sigma = E\{(X-\mu)(X-\mu)'\} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$$

Escoufier (1973) a défini le coefficient de corrélation vectoriel ρV :

$$\rho V = \rho V(X^{(1)}, X^{(2)}) = \frac{\mathrm{tr}\,(\Sigma_{12}\Sigma_{21})}{\sqrt{\mathrm{tr}\,(\Sigma_{11}^2)\mathrm{tr}\,(\Sigma_{22}^2)}}$$

Les propriétés essentielles de ce coefficient sont les suivantes :

i) Si $p = q = 1$ alors $\rho V = \rho^2$ le carré du coefficient de corrélation simple.

ii) $0 \leq \rho V \leq 1$ et

1) $\rho V = 0$ si et seulement si $\Sigma_{12} = 0$

2) $\rho V = 1$ si $X^{(2)} = AX^{(1)} + b$ où A est une matrice $q \times p$ telle que $A'A = kI$ (k scalaire positif) et b un vecteur $q \times 1$ (A et b constants).

iii) Si A est une matrice $p \times p$ telle que $A'A = kI$ (k scalaire positif) alors $\rho V(AX^{(1)}, X^{(2)}) = \rho V(X^{(1)}, X^{(2)})$.

L'utilisation de ce coefficient en statistique multivarié est multiple (Robert et Escoufier 1976, Dambroise, Escoufier et Massotte 1987, Cléroux, Helbling et Ranger 1990).

En présence d'un échantillon X_1, X_2, ..., X_n, on définit le coefficient de corrélation vectorielle échantillonnale en remplaçant dans l'expression de ρV les paramètres par les estimateurs habituels pour obtenir :

$$RV = RV(X^{(1)}, X^{(2)}) = \frac{\operatorname{tr}(S_{12}S_{21})}{\sqrt{\operatorname{tr}(S_{11}^2)\operatorname{tr}(S_{22}^2)}}$$

où $S_{ij} = \dfrac{1}{n-1}\sum(X_\alpha^{(i)} - \overline{X}^{(i)})(X_\alpha^{(j)} - \overline{X}^{(j)})' \quad i,j = 1,2,$

$\overline{X}^{(i)}$ et $\overline{X}^{(j)}$ désignant les parties respectives i et j des vecteurs moyennes empiriques des $X_\alpha^{(i)}$ et $X_\alpha^{(j)}$ $(1 \leq \alpha \leq n)$.

En définissant les matrices :

$$Y_1 = (X_1^{(1)} - \overline{X}^{(1)}, X_2^{(1)} - \overline{X}^{(1)}, \ldots, X_n^{(1)} - \overline{X}^{(1)}) : p \times n$$
$$Y_2 = (X_1^{(2)} - \overline{X}^{(2)}, X_2^{(2)} - \overline{X}^{(2)}, \ldots, X_n^{(2)} - \overline{X}^{(2)}) : q \times n$$

et en utilisant la norme $\|E\| = \sqrt{\operatorname{tr} E'E}$, Robert et Escoufier (1976) ont montré que :

$$\sqrt{2}\sqrt{1 - RV(X^{(1)}, X^{(2)})} = \operatorname{dist}(Y_1, Y_2) \qquad (*)$$

où $\operatorname{dist}(Y_1, Y_2) = \left\| \dfrac{Y_1'Y_1}{\sqrt{\operatorname{tr}(Y_1'Y_1)^2}} - \dfrac{Y_2'Y_2}{\sqrt{\operatorname{tr}(Y_2'Y_2)^2}} \right\|$

Ainsi la proximité entre les deux matrices de données Y_1 et Y_2 indique que la position relative des n points dans $\mathfrak{R}^p$ et dans $\mathfrak{R}^q$ est semblable.

2.3 La méthode basée sur le RV

Soit $X = (X_1, X_2, \ldots, X_n) = \begin{pmatrix} X_1^{(1)} X_2^{(1)} \ldots X_n^{(1)} \\ X_1^{(2)} X_2^{(2)} \ldots X_n^{(2)} \end{pmatrix} = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$, une matrice de données formée par n vecteurs individus de dimensions $p + q$.

Supposons qu'une partie du premier groupe de variables du vecteur X_n contienne des données manquantes, on notera ce vecteur inconnu x. Ainsi on obtient $X_n' = (x', y', z')$, y étant la partie complète du premier groupe de variables.

La méthode que nous proposons, consiste à remplacer x par le vecteur qui minimise dist (Y_1, Y_2) c-à-d, d'après la formule $(*)$ du paragraphe 2.2, par celui qui maximise $RV(X^{(1)}, X^{(2)})$.

Si l'indice i accolé à la matrice de variance-covariance S et au vecteur moyenne $\overline{X}$ indique que les i premiers individus ont participé à l'évaluation de cet estimateur, on a :

$$S_n = \frac{n-2}{n-1}\, S_{n-1} + \frac{1}{n}\,(X_n - \overline{X}_{n-1})(X_n - \overline{X}_{n-1})'$$

Posons $\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \frac{1}{\sqrt{n}}\,(X_n - \overline{X}_{n-1})$ et $\frac{n-2}{n-1}\, S_{n-1} = \left(\begin{array}{ccc} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ \hline A_{31} & A_{32} & A_{33} \end{array}\right)$

On a alors $S_n = \left(\begin{array}{ccc} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ \hline A_{31} & A_{32} & A_{33} \end{array}\right) + \left(\begin{array}{ccc} uu' & uv' & uw' \\ vu' & vv' & vw' \\ \hline wu' & wv' & ww' \end{array}\right)$

et donc en gardant les notations du paragraphe 2.2 :

$$S_{11} = \begin{pmatrix} A_{11} + uu' & A_{12} + uv' \\ A_{21} + vu' & A_{22} + vv' \end{pmatrix}, \quad S_{12} = \begin{pmatrix} A_{13} + uw' \\ A_{23} + vw' \end{pmatrix}, \quad S_{22} = A_{33} + ww'.$$

Le coefficient $RV(X^{(1)}, X^{(2)})$ dépend de x, qui est inconnu, par l'intermédiaire de u. On le notera $RV(u)$ et l'on doit donc maximiser $RV(u)$ en fonction de u. On obtient après des manipulations algébriques :

$$RV(u) = \frac{2w'A_{31}u + w'wu'u + \alpha}{\sqrt{\gamma}\sqrt{(u'u)^2 + 2u'(A_{11} + Iv'v)u + 4v'A_{21}u + \beta}}$$

avec
$$\begin{aligned} \alpha &= \operatorname{tr}\{A_{13}A_{31} + (A_{23} + vw')(A_{32} + wv')\} \\ &= \operatorname{tr}(A_{13}A_{31} + A_{32}A_{23}) + 2w'A_{32}v + w'wv'v \\ \beta &= \operatorname{tr}(A_{11}^2 + 2A_{12}A_{21} + A_{22}^2) + 2v'A_{22}v + (v'v)^2 \\ \gamma &= \operatorname{tr}\{(A_{33} + ww')^2\} = \operatorname{tr}(A_{33}^2) + 2w'A_{33}w + (w'w)^2 \end{aligned}$$

Le traitement de données artificielles (5 individus et 4 variables) permet de visualiser comment la méthode RV utilise la structure de Y_2 pour estimer la valeur

manquante de Y_1.

$$X = \left(\begin{array}{ccccc} 1 & 4 & 6 & 5 & \\ 3 & 4 & 3 & 1 & -1 \\ \hline 1 & 4 & 8 & 5 & 7 \\ 5 & 6 & 5 & 3 & 1 \end{array}\right)$$

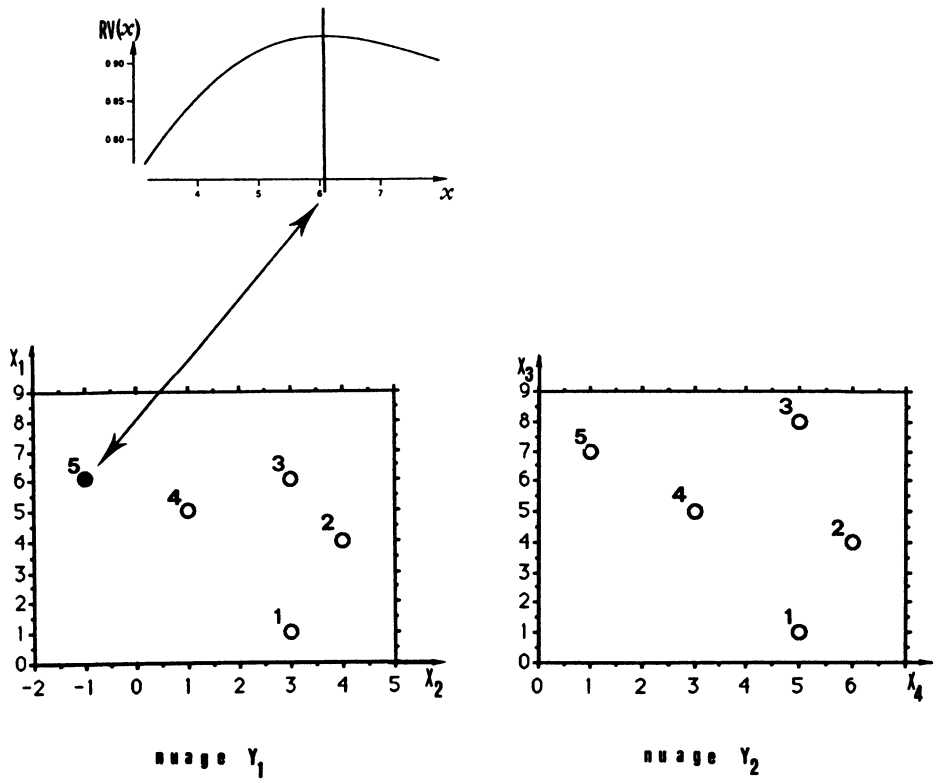
une valeur manque pour le dernier individu

Dans ce cas :

$$RV(u) = \frac{4.062u^2 + 11.739u + 41.563}{9.076\sqrt{u^4 + 12.625u^2 + 3.354u + 28.752}}$$

avec $u = (x - 4)/\sqrt{5}$

et $u_{\max} = 0.9257$, $x_{\max} = 6.07$, $RV_{\max} = 0.935$.



Remarques :

1) Si r vecteurs individus ont des données manquantes, on traite séparément et séquentiellement chacun des r vecteurs incomplets avec les $n - r$ vecteurs complets.

2) Si la séparation des données en deux groupes de variables n'est pas imposée par le contexte, on peut choisir une séparation de sorte que les données manquantes n'apparaissent que dans le premier groupe et éventuellement même afin que tout le vecteur manquant représente le premier groupe. Dés lors les formules se simplifient

$$S_n = \begin{pmatrix} A_{11} + uu' & A_{13} + uw' \\ A_{31} + wu' & A_{33} + ww' \end{pmatrix}$$

et

$$RV(u) = \frac{2w'A_{31}u + w'wu'u + \alpha}{\sqrt{\gamma}\sqrt{(u'u)^2 + 2u'(A_{11})u + \beta}}$$

avec $\alpha = \mathrm{tr}\,(A_{13}A_{31})$, $\beta = \mathrm{tr}\,(A_{11}^2)$ et $\gamma = \mathrm{tr}\,(A_{33}^2) + 2w'A_{33}w + (w'w)^2$

Si par contre la séparation est imposée à priori et que, malheureusement des données manquent dans les deux groupes de variables, le principe de maximisation de RV est toujours applicable. Seule l'expression du RV en fonction des parties manquantes est plus compliquée.

3) Dans le cas particulier des données bivariées ($p = q = 1$ et $\overline{X}_{n-1} = (c_1, c_3)$), $RV(u)$ devient :

$$RV(u) = \frac{(wu + A_{13})^2}{(w^2 + A_{33})(u^2 + A_{11})}$$

et par la propriété i) de ρV, est égal au carré de la corrélation simple.

La maximisation de cette fonction fournit l'estimation :

$$\widehat{x} = c_1 + \frac{A_{11}}{A_{13}}\,(z - c_3),$$

alors que la méthode de Buck estime par :

$$\widehat{x} = c_1 + \frac{A_{13}}{A_{33}}\,(z - c_3).$$

La méthode RV utilise donc la régression de z sur x et la méthode de Buck celle de x sur z.

4) Les propriétés ii) et iii) permettent d'assurer l'invariance de la méthode RV pour des transformations orthogonales.

3. Exemple et comparaison

Nous allons traiter les données «Têtes» recueillies par Frets et traitées dans le livre de Mardia, Kent et Bibby (1979). Elles forment une matrice de données X de 25 individus et 4 variables (2 groupes de 2 variables chacun). Nous y décrétons de manière aléatoire 4 valeurs manquantes : $X(23,1)$, $X(24,2)$, $X(25,1)$ et $X(25,2)$.

Pour l'approche *estimation directe* des données manquantes, nous comparerons la méthode RV avec quatre méthodes utilisant chacune le maximum d'information possible. La première méthode estime la donnée manquante par la moyenne (MEAN). Les trois suivantes sont basées sur la régression : régression multiple sur toutes les variables (REGR), régression simple sur la variable la plus corrélée (SINGLE) et régression pas-à-pas (STEP).

Pour l'approche *estimation des paramètres*, nous comparerons l'estimation de la matrice des corrélations après les estimations des valeurs manquantes des cinq méthodes précédentes et celle de trois nouvelles méthodes. La première est basée sur l'algorithme EM (ML), la seconde utilise toute l'information disponible pour le calcul de la matrice des corrélations (ALLVALUE) et la dernière ne prend en compte que les données complètes (COMPLETE).

Hormis la méthode RV, tous les résultats ont été obtenus avec BMDP. On peut donc se référer au manuel BMDP pour une description précise de ces méthodes (les abréviations entre parenthèses correspondent à la terminologie BMDP).

La comparaison se fera à l'aide de deux critères définis par Gleason et Staelin en 1975 : le premier Q_α représente une distance entre la valeur réelle et la valeur imputée, le second D_α représente une distance entre la corrélation réelle et la corrélation estimée.

$$\text{approche } \textit{estimation directe} : \quad Q_\alpha = \sqrt{\sum \frac{(X_{ij}^{(\alpha)} - X_{ij})^2}{\sigma_j^2 np\pi}}$$

$$\text{approche } \textit{estimation des paramètres} : \quad D_\alpha = \sqrt{\sum \frac{(R_{ij}^{(\alpha)} - R_{ij})^2}{p(p-1)}}$$

Dans ces formules, p est le nombre de variables, n le nombre d'individus, π le pourcentage de données manquantes, R_{ij} (resp. X_{ij}) la matrice des corrélations (resp. des données réelles), $R_{ij}^{(\alpha)}$ (resp. $X_{ij}^{(\alpha)}$) la matrice des corrélations (resp. des données) obtenue par la méthode α et σ_j^2 la variance réelle de la variable j.

Le tableau ci-après permet de faire les observations suivantes :

1) La méthode RV donne de bons résultats qui peuvent être expliqués par une valeur de RV assez élevée sur les données réelles ($RV = 0.5998$).

2) Le pourcentage de données manquantes dans notre exemple est faible ($\pi = 4\%$). Malgré cela la méthode MEAN est bien moins performante que les méthodes de type Buck (REGR, SINGLE, STEP). L'importance de la corrélation entre les variables étant un critère d'optimalité des méthodes de type Buck, leur

Méthode	D_α	Q_α
RV	0.01034	0.73959
MEAN	0.04280	1.56021
REGR	0.01043	1.11445
SINGLE	0.00896	1.21326
STEP	0.01053	1.12060
ML	0.01172	
ALLVALUE	0.02398	
COMPLETE	0.02208	

bon comportement ici n'a rien d'étonnant : en effet on a max $R_{ij} = 0.8392$ et min $R_{ij} = 0.6932$.

3) Il est surprenant de constater que les méthodes spécifiques d'estimation de paramètres ont un D_α moins bon que celles qui estiment les paramètres après imputation. Peut-on y voir une condamnation de la première approche ?

4) Pour voir l'influence de l'hypothèse de répartition au hasard des données manquantes, on a créé des données non DMCH (la donnée est manquante si $X_1 > 200$ ou $X_2 > 160$). Le coefficient Q_α augmente d'un facteur allant de 1.2 à 1.8 par rapport aux données DMCH mais la classification des méthodes reste identique : méthode RV en tête suivi des méthodes de type régression puis de la méthode basée sur la moyenne. Ainsi la méthode RV a montré sur cet exemple, qu'elle résistait mieux que les autres méthodes à la violation de cette hypothèse.

Le traitement d'autres données réelles a confirmé ces observations et a corroboré la bonne qualité d'imputation de la méthode RV.

Bibliographie

[1] ABRAHAM B. (1981) – Missing observations in time series, *Communications in Statistics : theory*, 10, 1643-1653.

[2] ANDERSON R.L.(1946) – Missing-plot techniques, *Biometrics*, 2, 41-47.

[3] BEALE E.M.L et LITTLE R.J.A. (1975) – Missing values in multivariate analysis, *Journal of Royal Statistical Society, Series B*, 37, 129-145.

[4] BOYLES R.A. (1983) – On the convergence of the EM algorithm, *Journal of Royal Statistical Society, Series B*, 45, 47-50.

[5] BUCK S.F. (1960) – A method of estimation of missing values in multivariate data suitable for use with an electronic computer, *Journal of Royal Statistical Society, Series B*, 22, 302-306.

[6] CLEROUX R., HELBLING J.-M. ET RANGER N.(1990) – Détection d'ensembles de données aberrantes en analyse des données multivariées, *Revue de Statistique Appliquée*, XXXVIII, 5-21.

[7] CURRY J. et KIM J.O. (1977) – The treatment of missing data in multivariate analysis, *Sociological Methods and Research*, 6, 215-240.

[8] DAMBROISE E., ESCOUFIER Y., MASSOTTE P. (1987) – Application de l'analyse de données à l'élaboration de mini-sondages d'opinion, *Revue de Statistique Appliquée*, XXXV, 9-24

[9] DEAR R.E.A. (1959) – A principal-component missing data method for multiple regression models. *System Development Corporation, Technical Report* SP-86.

[10] DEMPSTER A.P., LAIRD N.M. et RUBIN D.B. (1977) – Maximum likelihood from incomplete data via the EM algorithm, *Journal of Royal Statistical Society, Series B*, 39, 1-38.

[11] DER MEGREDITCHIAN G. (1988) – Problèmes engendrés par les données manquantes dans la pratique statistique, *Note de Travail de l'Etablissement d'Etudes et de Recherches Météorologiques*, 208.

[12] DODGE Y. (1985) – Analysis of experiments with missing data, *Wiley.*

[13] ESCOUFIER Y. (1973) – Le traitement des variables vectorielles, *Biometrics*, 29, 751-760.

[14] FORD B.L. (1983) – An overview of hot-deck procedures, dans *Incomplete Data in Sample Surveys, vol II : (W.G. Madow, I Olkin et D.B. Rubin, Eds), Academic Press.*

[15] FRANE J.W. (1976) – Some simple procedures for handling missing data in multivariate analysis, *Psychometrika*, 41, 409-415.

[16] GLEASON T.L. et STAELIN R. (1975) – Proposal for handling missing data, *Psychometrika*, 4, 229-252.

[17] GROSBRAS J.M. (1987) – Les réponses manquantes, *dans Les sondages édité par Droesbeke F, Fichet B. et Tassi P., Economica.*

[18] HAITOVSKY Y. (1968) – Missing data in regression analysis, *Journal of Royal Statistical Society, Series B*, 30, 67-82.

[19] HETJAN D.F.et RUBIN D.B. (1990) – Inference from coarse data via multiple imputation with application to age heaping, *Journal of the American Statistical Association*, 85, 304-314.

[20] HERZOG T.N. et RUBIN D.B. (1983) – Using multiple imputations to handle nonresponse in sample surveys, dans *Incomplete Data in Sample Surveys, vol II : (W.G. Madow, I Olkin et D.B. Rubin, Eds), Academic Press.*

[21] JONES R.H. (1980) – Maximum likelihood fitting of ARMA models to time series with missing observations, *Technometrics*, 22, 389-395.

[22] KAISER J. (1990) – The robustness of regression and substitution by mean methods in handling missing values, *présenté au* $22^{ème}$ *congrès de l'ASU à Tours.*

[23] KARIYA T, KRISHNAIAH P.R. et RAO C.R (1983) – Inference on parameters of multivariate normal populations when some data is missing, *Developments in Statistics*, 4, 137-184.

[24] LITTLE R.J.A. (1979) – Maximum likelihood inference for multiple regression with missing values : a simulation study, *Journal of Royal Statistical Society, Series B*, 41, 76-87.

[25] LITTLE R.J.A. (1986) – Survey nonresponse adjustement for estimates of means, *International Statistical Review*, 54, 139-157.

[26] LITTLE R.J.A. (1988) – A test of missing completely at random for multivariate data with missing values, *Journal of the American Statistical Association*, 83, 1198-1202.

[27] LITTLE R.J.A. et RUBIN D.B. (1987) – Statistical analysis with missing data, *Wiley*.

[28] MADOW W.G., OLKIN I. et RUBIN D.B. (eds) (1983) *Incomplete Data in Sample Surveys, 3 volumes, Academic Press.*

[29] MARDIA K.V., KENT J.T. et BIBBY J.M. (1979) – Multivariate analysis, *London, Academic Press.*

[30] MURRAY L.W. (1986) - Estimation of missing cells in randomized block and latin square designs, *The American Statistician*, 40, 289-293.

[31] ROBERT P. et ESCOUFIER Y (1976) – A unifying tool for linear multivariate statistical methods : the *RV*-coefficient, *Applied Statistics*, 25, 257-265.

[32] RUBIN D.B. (1972) – A non-iterative algorithm for least squares estimation of missing values in any analysis of variance design, *Journal of Royal Statistical Society, Series C*, 21, 136-141.

[33] RUBIN D.B. (1976) – Inference and missing data, *Biometrika*, 63, 581-592.

[34] RUBIN D.B. (1987) – Multiple imputation for nonresponse in survey, *Wiley*.

[35] RUBIN D.B. et SCHENKER N. (1986) – Multiple imputation for interval estimation from simple random samples with ignorable nonresponse, *Journal of the American Statistical Association*, 81, 366-374.

[36] SEBER G.A.F.(1984) – Multivariate observations, *Wiley*, 422-423

[37] SIMON G.A. et SIMONOFF J.S. (1986) – Diagnostic plots for missing data in least squares regression, *Journal of the American Statistical Association*, 81, 501-509.

[38] SRIVASTAVA M.S. (1985) – Multivariate data with missing observations, *Commun. Statis. – Theor. Meth.* , 14, 775-792.

[39] TIMM N.H. (1970) – The estimation of variance-covariance and correlation matrices from incomplete data, *Psychometrika*, 35, 417-435.