REVUE DE STATISTIQUE APPLIQUÉE

GUY DER MEGREDITCHIAN

Problèmes engendrés par les données manquantes dans la pratique statistique

Revue de statistique appliquée, tome 40, n° 1 (1992), p. 7-22

<http://www.numdam.org/item?id=RSA_1992__40_1_7_0>

Rev. Statistique Appliquée, 1992, **XXXIX** (1), 7-22

PROBLÈMES ENGENDRÉS PAR LES DONNÉES MANQUANTES DANS LA PRATIQUE STATISTIQUE

Guy Der MEGREDITCHIAN † [1]

METEO-FRANCE
2 Avenue Rapp, 75340 PARIS CEDEX 07

RÉSUMÉ

Les fichiers perturbés par des données manquantes ne peuvent être traités avec les algorithmes prévus pour les fichiers complets. L'application formelle des formules classiques entraine l'apparition de diverses incohérences tant pour les estimations des paramètres individuels, (variances, covariances, corrélations) que pour la signification des matrices formées à partir de ces paramètres. Une difficulté analogue se manifeste pour le calcul des coefficients d'autocovariance et d'autocorrélation même évalués sur un fichier complet, car le décalage temporel nous replace dans le cadre d'un fichier incomplet. Enfin dans la routine prévisionnelle les valeurs manquantes des prédicteurs entravent la réalisation de la prévision. Les artifices habituellement proposés ne sont guère satisfaisants.

Une réponse claire est apportée aux problèmes spécifiques que soulèvent les données manquantes, de sorte que l'on peut envisager un traitement statistique correct des fichiers incomplets.

Mots-clés : *Autocovariance, Covariance, Données Manquantes, Estimation, Régression, Variance.*

1. Introduction

Le praticien est souvent confronté au problème suivant. Il dispose de formules appropriées pour le calcul des paramètres statistiques à partir d'un fichier de données, mais celles-ci ne sont valables, rigoureusement parlant, que dans le cas où les fichiers sont complets. Or en réalité il est fréquent que des données manquantes perturbent les fichiers disponibles et les formules classiques ne sont plus applicables. Que faire en pareil cas ? Que se passe-t-il si malgré tout on applique formellement les formules classiques ?

[1] Article soumis en Mai 1988 et révisé en novembre 1990 suivant les indications du comité de rédaction par Patricia POTTIER en hommage à l'auteur décédé en Février 1990.

Pour bien comprendre le type de problèmes qui se posent considérons le mini fichier suivant :

x_1	$x_1[1]$	$x_1[2]$	$x_1[3]$	
x_2		$x_2[2]$	$x_2[3]$	$x_2[4]$

(1)

Introduisons les notations :

$\mathcal{T}$ l'ensemble des dates d'observation des variables x_1 et x_2 :

$$\mathcal{T} = \{1, 2, 3, 4\} \; ; \quad T = card[\mathcal{T}] = 4$$

$\mathcal{T}_1$ l'ensemble des dates d'observation de la variable x_1 :

$$\mathcal{T}_1 = \{1, 2, 3\} \; ; \quad T_1 = card[\mathcal{T}_1] = 3$$

$\mathcal{T}_2$ l'ensemble des dates d'observation de la variable x_2 :

$$\mathcal{T}_2 = \{2, 3, 4\} \; ; \quad T_2 = card[\mathcal{T}_2] = 3$$

$\mathcal{T}_{12}$ l'ensemble des dates où l'on a observé simultanément x_1 et x_2 :

$$\mathcal{T}_{12} = \{2, 3\} \; ; \quad T_{12} = card[\mathcal{T}_{12}] = 2$$

Posons encore pour simplifier les calculs :

$$\left\{ \begin{array}{lcl} x_1[1] & = & a \\ x_1[2] & = & \text{-}pa \\ x_1[3] & = & pa \\ x_2[2] & = & \text{-}qb \\ x_2[3] & = & \text{-}qb \\ x_2[4] & = & b \end{array} \right.$$

Calculons maintenant le coefficient de corrélation :

$$\hat{r}[x_1, x_2] = \frac{\hat{v}_{x_1 x_2}}{\hat{\sigma}_{[x_1]} \hat{\sigma}_{[x_2]}} \tag{2}$$

Notons tout d'abord que formellement 64 valeurs différentes peuvent être obtenues pour cette estimation. En effet les valeurs moyennes $\bar{x}_i$ et les variances $\hat{\sigma}^2_{[x_i]}$ peuvent être calculées soit sur tout le fichier disponible ($\mathcal{T}_1$ ou $\mathcal{T}_2$), soit sur l'intersection $\mathcal{T}_{12}$. Cela devient évident, si l'on met la formule (2) sous la forme :

$$r[x_1, x_2] = \frac{\frac{1}{N} \sum_{t \in \mathcal{T}_{[1,2]}} [x_1(t) - \frac{1}{N_1} \sum_{t \in \mathcal{T}_{[1]}} x_1(t)][x_2(t) - \frac{1}{N_2} \sum_{t \in \mathcal{T}_{[2]}} x_2(t)]}{\sqrt{\frac{1}{N_3} \sum_{t \in \mathcal{T}_{[3]}} [x_1(t) - \frac{1}{N_4} \sum_{t \in \mathcal{T}_{[4]}} x_1(t)]^2 \frac{1}{N_5} \sum_{t \in \mathcal{T}_{[5]}} [x_2(t) - \frac{1}{N_6} \sum_{t \in \mathcal{T}_{[6]}} x_2(t)]^2}} \tag{3}$$

Dans la formule (3) six sommations sont à définir, c'est pourquoi nous les avons noté $t \in \mathcal{T}_{[K]}, (K = 1 \text{ à } 6)$. A chaque fois le choix de $\mathcal{T}_{[K]}$ peut être fait entre $\mathcal{T}_1$, $\mathcal{T}_2$ ou $\mathcal{T}_{12}$. Conformément à ce choix on doit également choisir la norme N_K qui doit être respectivement T_1, T_2 ou T_{12}.

Cela nous conduit formellement à $2^6 = 64$ variantes différentes (dont la plupart sont évidemment déraisonnables) pour calculer une estimation $\tilde{r}$ de r. En outre pour chaque variante particulière se pose le problème du choix adéquat de la norme N. On adoptera un principe apparemment rationnel pour choisir l'une de ces 64 variantes, celui de l'utilisation exhaustive de toute l'information disponible.

On choisira ainsi :

$$\left\{\begin{array}{llllll} \mathcal{T}_{[1]} & = & \mathcal{T}_1 & ; & N_1 & = & T_1 \\ \mathcal{T}_{[2]} & = & \mathcal{T}_2 & ; & N_2 & = & T_2 \\ \mathcal{T}_{[3]} & = & \mathcal{T}_1 & ; & N_3 & = & T_1 - 1 \\ \mathcal{T}_{[4]} & = & \mathcal{T}_1 & ; & N_4 & = & T_1 \\ \mathcal{T}_{[5]} & = & \mathcal{T}_2 & ; & N_5 & = & T_2 - 1 \\ \mathcal{T}_{[6]} & = & \mathcal{T}_2 & ; & N_6 & = & T_2. \end{array}\right.$$

Pour $\mathcal{T}_{[1,2]}$, l'unique possibilité est évidemment $\mathcal{T}_{[1,2]} = \mathcal{T}_{12}$, de sorte que l'on peut prendre $N = T_{12} - 1$.

On obtient successivement :

$\bar{x}_1 = a/3$

$\bar{x}_2 = b/3$

$\tilde{v}_{12} = \frac{1}{1}[-pa - a/3][-qb - b/3] + [pa - a/3][qb - b/3] = 2ab[pq + 1/9]$

$\tilde{\sigma}_1^2 = \frac{1}{2}[(a - a/3)^2 + (-pa - a/3)^2 + (pa - a/3)^2] = a^2[p^2 + 1/3].$

De même :

$\tilde{\sigma}_2^2 = b^2[q^2 + 1/3].$

On obtient ainsi pour le coefficient de corrélation empirique :

$$\tilde{r}_{12} = \frac{\tilde{v}_{x_1 x_2}}{\tilde{\sigma}_{[x_1]}\tilde{\sigma}_{[x_2]}} = \frac{2ab[pq + 1/9]}{\sqrt{a^2(p^2 + 1/3)b^2(q^2 + 1/3)}} = \frac{2[pq + 1/9]}{\sqrt{(p^2 + 1/3)(q^2 + 1/3)}} \tag{4}$$

Les valeurs p et q arbitraires, on a toute latitude pour les choisir de telle sorte que la valeur de $\tilde{r}_{12}$ soit incohérente ($\prec -1$ ou $\succ 1$). Ainsi on peut choisir p et q de telle sorte que :

$$\tilde{r}_{12} \prec -1 \quad ou \quad \tilde{r}_{12} \succ 1$$

Cela est vérifié aisément. En particulier si l'on pose $p = 1$, $q = -1$ on obtient :

$$\tilde{r}_{12} = -4/3$$

et si l'on prend $p = 1$, $q = 1$:

$$\tilde{r}_{12} = 5/3$$

Plus généralement, si l'on choisit $p = q$, cela donne :

$$\tilde{r}_{12} = -1 + \frac{p^2 - 1/9}{p^2 + 1/3}$$

et $\tilde{r}_{12} \succ 1$, si $|p| \succ 1/3$.

De même, si l'on pose $p = -q$ cela donne :

$$\tilde{r}_{12} = -1 - \frac{p^2 - 5/9}{p^2 + 1/3}$$

et $\tilde{r}_{12} \prec -1$, si $|p| \succ \sqrt{5}/3$.

L'exemple présenté illustre bien le fait que les données manquantes perturbent le calcul des paramètres statistiques et que les formules classiques ne sont plus valables. Il faut donc revoir le problème de manière à établir des estimations sans biais de chaque paramètre pris individuellement. De plus il faut élucider la cohérence du procédé de calcul de plusieurs paramètres empiriques.

Evidemment il n'y a aucune difficulté de calcul de r_{12}, si l'on décide de travailler uniquement sur le fichier intersection $\mathcal{T}_{12} = \mathcal{T}_1 \cap \mathcal{T}_2$ car dans de cas, tout se passe du point de vue numérique, comme si l'on avait un fichier complet.

Cela ne résoud pas le problème pour autant, car d'une part il est dommage de ne pas utiliser intégralement l'information disponible (pourquoi rejeter le fichier $\mathcal{T}_1 \setminus \mathcal{T}_2 \cup \mathcal{T}_2 \setminus \mathcal{T}_1$) et d'autre part cette procédure n'est plus valable pour le calcul d'une matrice de corrélation car très souvent l'intersection

$$\mathcal{T}_{1,2,\ldots,n} = \bigcap_{i=1}^{n} \mathcal{T}_i$$

est vide.

2. Estimation non biaisée du coefficient de covariance

Nous adopterons le principe raisonnable d'utilisation maximum de l'information disponible. Autrement dit le coefficient de covariance sera recherché sous la forme :

$$\tilde{v}_{ij} = \frac{1}{N_{ij}} \sum_{t \in \mathcal{T}_{ij}} [x_i(t) - \bar{x}_i][x_j(t) - \bar{x}_j], \tag{5}$$

où

$$\left\{ \begin{array}{lcl} \bar{x}_i & = & \frac{1}{N_i} \sum_{t \in \mathcal{T}_i} x_i(t) \\ \bar{x}_j & = & \frac{1}{N_j} \sum_{t \in \mathcal{T}_j} x_j(t) \end{array} \right.$$

Le paramètre N_{ij} dépendra évidemment de T_i, T_j et T_{ij} :

$$N_{ij} = N[T_i, T_j, T_{ij}]$$

et nous voulons l'estimer à partir de la condition de non biais :

$$E[\tilde{v}_{ij}] = v_{ij}$$

Soit :

$$\left\{ \begin{array}{lcl} \mathcal{T}_i & = & \{t_1, ..., t_s, ..., t_{T_i}\} \\ \mathcal{T}_j & = & \{\tau_1, ..., \tau_s, ..., \tau_{T_j}\} \\ \mathcal{T}_{ij} & = & \{\theta_1, ..., \theta_s, ..., \theta_{T_{ij}}\} \\ X_i & = & \{x_i(1), ..., x_i(t), ..., x_i(T)\} \\ X_j & = & \{x_j(1), ..., x_j(t), ..., x_j(T)\} \end{array} \right.$$

On considère que les variables x_i et x_j sont centrées, ce qui ne restreint pas la généralité puisque la formule (5) reste inchangée si on remplace x_i par $x_i - m_i$ et x_j par $x_j - m_j$ (en notant $m_i = E[x_i]$ et $m_j = E[x_j]$).

De plus on supposera que le couple $\{x_i, x_j\}$ suit une loi normale, autrement dit :

$$\{x_i, x_j\} \in \mathcal{N}\left[0, V_{x_i x_j}\right] \quad ; \quad V_{x_i x_j} = \begin{pmatrix} \sigma_i^2 & v_{ij} \\ v_{ji} & \sigma_j^2 \end{pmatrix}.$$

L'échantillon global est constitué d'observations indépendantes, de sorte que le couple de vecteurs $\{X_i, X_j\}$ suit une loi normale de paramètres :

$$\mathcal{M} = \{0, 0\} \quad ; \quad V_{X_i X_j} = \begin{pmatrix} \sigma_i^2 I_T & v_{ij} I_T \\ v_{ji} I_T & \sigma_j^2 I_T \end{pmatrix} .$$

Etudions la variable aléatoire ξ_{ij} :

$$\xi_{ij} = \sum_{t \in \mathcal{T}_{ij}} [x_i(t) - \bar{x}_i][x_j(t) - \bar{x}_j] .$$

Transformons l'expression de ξ_{ij} en ouvrant les parenthèses :

$$\xi_{ij} = \sum_{t \in \mathcal{T}_{ij}} x_i(t) x_j(t) - \bar{x}_i \sum_{t \in \mathcal{T}_{ij}} x_j(t) - \bar{x}_j \sum_{t \in \mathcal{T}_{ij}} x_i(t) + \sum_{t \in \mathcal{T}_{ij}} \bar{x}_i \bar{x}_j .$$

Désignons par $e_T(i)$ le vecteur $T \times 1$, dont toutes les composantes sont nulles sauf la $i^{ème}$ qui vaut 1, et posons :

$$E_i = \sum_{s=1}^{T_i} e_T(t_s) \quad ; \quad E_j = \sum_{s=1}^{T_j} e_T(\tau_s) \quad ; \quad E_{ij} = \sum_{s=1}^{T_{ij}} e_T(\theta_s) \quad ;$$

$$A_{ij} = (e_T(\theta_1), ..., e_T(\theta_s), ..., e_T(\theta_{T_{ij}}) \quad ; \quad A_{ij} \textit{ est une matrice } T \times T_{ij} \quad .$$

On obtient alors les résultats suivants :

Lemme 1

$$\left\{ \begin{array}{lcccl} \bar{x}_i & = & \frac{1}{T_i} \sum_{t \in \mathcal{T}_i} x_i(t) & = & \frac{1}{T_i} E_i' X_i \\ \bar{x}_j & = & \frac{1}{T_j} \sum_{t \in \mathcal{T}_j} x_j(t) & = & \frac{1}{T_j} E_j' X_j \ . \end{array} \right.$$

Lemme 2

$$\sum_{t \in \mathcal{T}_{ij}} x_i(t) = E_{ij}' X_i \quad ; \quad \sum_{t \in \mathcal{T}_{ij}} x_j(t) = E_{ij}' X_j \ .$$

Lemme 3

$$\sum_{t \in \mathcal{T}_{ij}} x_i(t) x_j(t) = X_i' A_{ij} A_{ij}' X_j \ .$$

En effet :

$$\begin{array}{rcccccccc} X_i' A_{ij} & = & (X_i' e_T(\theta_1), & ..., & X_i' e_T(\theta_s), & ..., & X_i' e_T(T_{ij})) \\ & = & (x_i(\theta_1), & ..., & x_i(\theta_s), & ..., & x_i(T_{ij})) \end{array}$$

De même :

$X_j' A_{ij} = (x_j(\theta_1), ...x_j(\theta_s), ..., x_j(T_{ij}))$

Par conséquent :

$X_i' A_{ij} A_{ij}' X_j = \sum_{s=1}^{T_{ij}} x_i(\theta_s) x_j(\theta_s) = \sum_{t \in T_{ij}} x_i(t) x_j(t)$.

On obtient ainsi une forme bilinéaire pour exprimer ξ_{ij} :

$$\xi_{ij} = X_i' [A_{ij} A_{ij}' - \frac{1}{T_i} E_i E_{ij}' - \frac{1}{T_j} E_{ij} E_j' + \frac{T_{ij}}{T_i T_j} E_i E_j'] X_j \ .$$

En posant :

$B_{ij} = A_{ij} A_{ij}' - \frac{1}{T_i} E_i E_{ij}' - \frac{1}{T_j} E_{ij} E_j' + \frac{T_{ij}}{T_i T_j} E_i E_j'$,

nous pouvons mettre ξ_{ij} sous la forme générale :

$$\xi_{ij} = X_i' B_{ij} X_j \ .$$

On a alors en désignant par Tr la Trace :

$$E[\xi_{ij}] = E[Tr(\xi_{ij})] = E[Tr(X_i' B_{ij} X_j)] = E[Tr(X_j X_i' B_{ij})] = Tr(V_{X_i X_j} B_{ij}) = v_{ij} Tr(B_{ij})$$

$$\begin{aligned} E[\xi_{ij}] &= v_{ij}[\ Tr(A_{ij}A_{ij}') - \tfrac{1}{T_i}Tr(E_iE_{ij}') - \tfrac{1}{T_j}Tr(E_{ij}E_j') + \tfrac{T_{ij}}{T_iT_j}Tr(E_iE_j')] \\ &= v_{ij}[\ I - \tfrac{1}{T_i}.II - \tfrac{1}{T_j}.III + \tfrac{T_{ij}}{T_iT_j}.IV] \end{aligned}$$

Le terme I se calcule aisément :

$I = Tr(A_{ij}A_{ij}') = Tr(A_{ij}'A_{ij}) = Tr\left((..., e_T'(\theta_s), ...)'(..., e_T(\theta_s))\right)$

$= Tr\left(\sum_{s=1}^{T_{ij}} e_T'(\theta_s) e_T(\theta_s)\right) = \sum_{s=1}^{T_{ij}} e_T'(\theta_s) e_T(\theta_s)) = \sum_{s=1}^{T_{ij}} 1 = T_{ij}$

De même nous avons directement pour les autres termes II, III, IV :

$$\begin{aligned} II &= Tr(E_iE_{ij}') = E_{ij}'E_i = T_{ij}\ ; \\ III &= Tr(E_{ij}E_j') = E_j'E_{ij} = T_{ij}\ ; \\ IV &= Tr(E_iE_j') = E_i'E_j = T_{ij}\ . \end{aligned}$$

Nous obtenons ainsi :

$$E[\xi_{ij}] = v_{ij}[T_{ij} - \frac{T_{ij}}{T_i} - \frac{T_{ij}}{T_j} + \frac{T_{ij}^2}{T_iT_j}] = v_{ij}[T_iT_j - T_i - T_j + T_{ij}]\frac{T_{ij}}{T_iT_j}\ .$$

L'estimation non biaisée sera :

$$\tilde{v}_{ij} = \frac{T_iT_j}{T_{ij}[T_iT_j - T_i - T_j + T_{ij}]} \sum_{t \in \mathcal{T}_{ij}} [x_i(t) - \bar{x}_i][x_j(t) - \bar{x}_j], \qquad (6)$$

Nous avons obtenu le coefficient de norme N_{ij} à partir de la condition de non biais :

$$N_{ij} = \frac{T_{ij}[T_iT_j - T_i - T_j + T_{ij}]}{T_iT_j} \qquad (7)$$

On remarquera que cette formule est valable dans tous les cas et aussi bien entendu quand il n'y a pas de données manquantes. Dans ce cas nous avons :

$$T_i = T_j = T_{ij} = T \quad et$$

$$N_{ij} = \frac{T[T^2 - T - T + T]}{T^2} = T - 1\ .$$

ce qui correspond bien à la norme de l'estimation non biaisée dans le cas où il n'y a pas de données manquantes.

3. Variance d'échantillonnage de l'estimation non biaisée du coefficient empirique de covariance

On peut, en utilisant les résultats classiques sur les formes bilinéaires et les formes quadratiques de variables gaussiennes, calculer la variance $\mathcal{D}(\tilde{v}_{ij})$ de $\tilde{v}_{ij} = \frac{\xi_{ij}}{N_{ij}}$.

On a :

$$\mathcal{D}(\tilde{v}_{ij}) = \mathcal{D}(\frac{X_i' B_{ij} X_j}{N_{ij}}) = \frac{2}{N_{ij}^2} Tr[(V_{X_j X_i} B_{ij})^2] = \frac{2v_{ij}^2}{N_{ij}^2} Tr[B_{ij}^2]$$

d'où l'on déduit, après un calcul fastidieux, mais sans difficultés :

$$\mathcal{D}(\tilde{v}_{ij}) = \frac{2v_{ij}}{T_{ij}(\theta_{ij} + T_i T_j)^2} [T_i^2 T_j^2 + 2T_i T_j \theta_{ij} + T_{ij} \theta_{ij}^2] \tag{8}$$

avec $\theta_{ij} = T_{ij} - (T_i + T_j)$.

S'il n'y a pas de données manquantes alors : $\theta_{ij} = -T$.

et l'expression de la variance devient : $\mathcal{D}(\tilde{v}_{ij}) = \frac{2v_{ij}}{T-1}$

4. Autocovariance et autocorrélation

Considérons la série $\mathcal{Z}$ des T observations de la variable Z :

$$\mathcal{Z} = Z(1), ..., Z(t), ..., Z(T) \ .$$

Pour calculer l'autocovariance de cette série

$$v_Z(\tau) = cov[Z(t), Z(t+\tau)] \ ,$$

on décale le fichier $\mathcal{Z}$ de τ unités, ce qui donne :

$$\begin{array}{ccccccccc} Z(1), & ..., & Z(\tau+1), & ..., & Z(t), & ..., & Z(T) & & \\ & & Z(1), & ..., & Z(t-\tau), & ..., & Z(T-\tau), & ..., & Z(T) \end{array}$$

Ramenons nous au modèle précédent en posant :

$$x_1(t) = Z(t) \ \ ; \ \ x_2(t) = Z(t+\tau) \ .$$

On a ainsi :

$$T_1 = T = T_2 \ \ ; \ \ T_{12} = T - \tau \ .$$

Le paramètre N figurant dans le calcul de l'estimation non biaisée de la covariance devient alors :

$$N = \frac{T_{12}[T_1T_2 - T_1 - T_2 + T_{12}]}{T_1T_2} = \frac{(T-\tau)[T^2 - T - \tau]}{T^2} .$$

La formule de calcul du coefficient d'autocovariance devient ainsi :

$$v_Z(\tau) = \frac{T^2}{(T-\tau)[T^2 - T - \tau]} \sum_{t=1}^{T-\tau} [Z(t) - \bar{Z}][Z(t+\tau) - \bar{Z}] , \qquad (9)$$

où $\bar{Z} = \frac{1}{T}\sum_{t=1}^{T} Z(t)$.
Le coefficient d'autocorrélation est alors par définition :

$$r_Z(\tau) = \frac{cov[Z(t), Z(t+\tau)]}{\sqrt{\mathcal{D}(Z(t))\mathcal{D}(Z(t+\tau))}} = \frac{v_Z(\tau)}{v_Z(0)} .$$

Nous obtenons ainsi l'expression suivante de $v_Z(0)$:

$$v_Z(0) = \frac{T^2}{T(T^2 - T)} \sum_{t=1}^{T} [Z(t) - \bar{Z}]^2 = \frac{1}{T-1} \sum_{t=1}^{T} [Z(t) - \bar{Z}]^2 ,$$

d'où nous tirons l'expression de $r_Z(\tau)$:

$$r_Z(\tau) = \frac{T^2(T-1)}{(T-\tau)(T^2 - T - \tau)} \frac{\sum_{t=1}^{T-\tau} [Z(t) - \bar{Z}][Z(t+\tau) - \bar{Z}]}{\sum_{t=1}^{T} [Z(t) - \bar{Z}]^2} . \qquad (10)$$

5. Recherche de la matrice de covariance "la plus proche" du tableau V

Après avoir calculé les estimations non biaisées du coefficient de covariance on peut former le tableau.

$$\tilde{V} = (\tilde{v}_{ij}) .$$

Toutefois $\tilde{V}$ peut ne pas appartenir à la famille des matrices de covariance. En effet par définition les matrices de covariance théorique V et empirique $\hat{V}$ (dans le cas des fichiers complets) sont des matrices définies non négatives :

$$V \in def(\succeq 0) \quad ; \quad \hat{V} \in def(\succeq 0) .$$

Pour la matrice de covariance empirique cela découle du fait que dans le cas où le fichier est complet on a la relation :

$$\hat{V} = \frac{1}{T} \mathcal{X}\mathcal{X}' ,$$

(où $\mathcal{X}$ est le tableau $(n \times T)$ des données centré); autrement dit $\hat{V}$ est une matrice de Gramm et l'on sait que les matrices de Gramm sont définies non négatives.

Pour le tableau $\tilde{V}$ obtenu à partir d'un fichier incomplet, perturbé par des données manquantes, la représentation n'est plus valable et par conséquent le tableau $\tilde{V}$ n'a aucune raison d'être une matrice définie non négative, d'où il ressort que $\tilde{V}$ n'appartient pas à la famille des matrices de covariance. C'est la raison pour laquelle toutes sortes d'incohérences peuvent se produire lors du calcul de la matrice de corrélation empirique. En effet rappelons que la condition :

$$r_{ij}^2 \preceq 1 \ ,$$

découle précisément du fait que la matrice de corrélation doit être une matrice définie positive. Les conditions nécessaires pour cela sont celles de Sylvestre suivant lesquelles il faut que les mineurs principaux de R soit non négatifs. En particulier cela donne :

$$\begin{vmatrix} 1 & r_{ij} \\ r_{ij} & 1 \end{vmatrix} \succeq 0 \ ,$$

d'où découle la relation $r_{ij}^2 \preceq 1$,

C'est pourquoi il est rationnel de choisir en qualité d'estimation de la matrice V une matrice $\hat{V}$ qui d'une part appartienne à la famille des matrices de covariance et d'autre part soit la plus proche du tableau $\tilde{V}$ au sens d'une métrique appropriée.

Définissons la distance $d^2(A, B)$ entre deux matrices symétriques A et B par la relation :

$$d^2(A, B) = Tr[A - B]^2 = \sum_{i,j=1}^{n} [a_{ij} - b_{ij}]^2 \ .$$

Soit $A = C\Lambda C'$ la décomposition usuelle de A suivant ses valeurs propres et ses vecteurs propres, C désignant la matrice orthogonale des vecteurs propres de A et Λ la matrice diagonale des valeurs propres.

Si l'on pose $F = C'BC$, auquel cas $B = CFC'$ (puisque C'C=CC'=I), on a :

$$A - B = C\Lambda C' - B = C(\Lambda - C'BC)C' = C(\Lambda - F)C'.$$

$$(A - B)^2 = C(\Lambda - F)'C'C(\Lambda - F)C' = C(\Lambda - F)^2C'.$$

$$Tr[(A - B)^2] = Tr[C(\Lambda - F)^2C'] = Tr[C'C(\Lambda - F)^2] = Tr[(\Lambda - F)^2] \ .$$

Désignant par $\lambda_i (i = 1, k)$ les valeurs propres positives de A, par $-|\lambda_j| (j = k + 1, n) (\lambda_j \preceq 0)$ les valeurs propres négatives de A, et par f_{ij} les éléments de F, on a :

$$Tr[(A - B)^2] = \sum_{i=1}^{k} (\lambda_i - f_{ii})^2 + \sum_{j=k+1}^{n} (-|\lambda_j| - f_{jj})^2) + \sum_{i,j,i \neq j} f_{ij}^2 \qquad (11)$$

Compte-tenu de ce que si B est définie ou semi définie positive, il en est de même de F, on déduit de (11) que la matrice $B = CFC'$ définie ou semi définie positive minimisant $Tr[(A-B)^2]$ est telle que F est diagonale, avec :

$$\begin{cases} f_{ii} = \lambda_i & si \quad 1 \preceq i \preceq k \\ f_{jj} = 0 & si \quad k+1 \preceq j \preceq n \end{cases}$$

On en déduit le théorème suivant, en prenant

$A = \tilde{V}$ et $B = \hat{V}$.

Théorème 1 *La matrice V de covariance la plus proche du tableau $\tilde{V}$ au sens de la métrique :*

$$d^2[\tilde{V}, V] = Tr[(\tilde{V} - V)^2],$$

est la matrice :

$$\hat{V} = C\Lambda^+ C',$$

où

- $C = C(\tilde{V})$ *est la matrice des vecteurs propres de* $\tilde{V}$ *et*
- $\Lambda^+ = \Lambda \begin{pmatrix} I_k & 0 \\ 0 & 0 \end{pmatrix}$ *s'obtient en remplaçant dans la matrice* $\Lambda = \Lambda(\tilde{V})$ *des valeurs propres de* $\tilde{V}$ *les valeurs propres négatives par O.*

6. Recalcul automatique des coefficients de régression quand certains prédicteurs sont manquants

Un grave inconvénient des données manquantes se manifeste lors de la réalisation en routine opérationnelle de la prévision statistique. En effet, alors même que l'on a élaboré une équation de régression linéaire multiple, on a avec des notations évidentes :

$$\tilde{y} = V_{yX} V_{XX}^{-1} [X - \mathcal{M}_X] + m_y = \sum_{i=1}^{n} a_i x_i + a_0, \tag{12}$$

où $\begin{cases} \mathcal{A} = V_{XX}^{-1} V_{Xy} & = \{a_1, ..., a_n\}, \\ a_0 = m_y + \mathcal{A}' \mathcal{M}_X. \end{cases}$

il s'avère souvent impossible d'utiliser (12) dans les conditions opérationnelles.

En effet si un ou plusieurs prédicteurs sont manquants la formule (12) est inexploitable. L'artifice habituellement utilisé en pareil cas consiste à remplacer le prédicteur manquant par sa valeur moyenne, ce qui généralement est loin d'être satisfaisant.

Supposons que le prédicteur manquant soit précisément x_k.

Il est évidemment impossible de prévoir à l'avance ou de stocker les coefficients correspondants pour tous les cas de figure possibles. Pourtant si l'on dispose d'une information complémentaire, le recalcul des coefficients de régression et de discrimination peut s'effectuer de façon très simple.

Ainsi supposons que l'on dispose au départ des renseignements suivants :

- 1) les n prédicteurs sélectionnés :

$$X = \{x_1, ..., x_i, ..., x_n\};$$

- 2) les coefficients de régression :

$$\mathcal{A} = \{a_1, ..., a_i, ..., a_n\};$$

- 3) la matrice inverse de variance-covariances :

$$V^{-1} = (v^{(i,j)}) = (V^{(1)}, ...V^{(i)}, ...V^{(n)}).$$

(où V a été calculé à partir de $\tilde{V}$ suivant le théorème donné au paragraphe 5..)

Nous avons noté $v^{(i,j)}$ l'élément (i, j) de la matrice inverse et $V^{(j)}$ la j-ième colonne de la matrice inverse :

$$V^{(j)} = \{v^{(1,j)}, ..., v^{(i,j)}, ..., v^{(n,j)}\}'.$$

Cette notation a été adoptée pour la différencier de la notation traditionnelle pour la matrice de variances-covariances :

$$V = (v_{ij}) = (V_1, ..., V_j, ..., V_n).$$

Ainsi les indices inférieurs concernent la matrice V et les indices supérieurs la matrice V^{-1}.

Supposons maintenant que le prédicteur x_k soit manquant à l'instant t. Nous disposons ainsi d'un vecteur à $n - 1$ dimensions que nous noterons :

$$X^{[k]} = \{x_1, ..., x_{k-1}, x_{k+1}, ..., x_n \ \} \ ,$$

ou encore,

$$Z \quad = \{z_1, ..., z_{k-1}, z_k \ \ , ..., z_{n-1}\} \ .$$

La notation $X^{[k]}$ correspond au fait que parmi les composantes du vecteur X on a enlevé la k-ième. L'intérêt d'utiliser une double notation $X^{[k]} \equiv Z$ est de ne pas avoir de saut dans la numérotation des indices.

Nous devons trouver la nouvelle équation de régression (ou de discrimination) avec les $n - 1$ variables disponibles,

$$\tilde{\tilde{y}} = b' X^{[k]} = b' Z = \sum_{i=1}^{n-1} b_i z_i$$

à partir de l'équation prévisionnelle (régression ou discrimination) correspondant à l'information complète :

$$\tilde{y} = b'X$$

Notre but est de recalculer le vecteur b à partir de l'information disponible, c'est-à-dire de la matrice V^{-1} et du vecteur $X^{[k]}$. En d'autres termes on recherche :

$$b = \psi[\mathcal{A}, V^{-1}].$$

Le résultat est concrétisé par le théorème suivant :

Théorème 2 *Le vecteur des coefficients de régression du vecteur $X^{[k]}$ à $n-1$ dimensions peut être exprimé par la relation :*

$$b = \mathcal{A}^{[k]} - \frac{a_k}{v^{(k,k)}} V^{[k](k)}. \tag{13}$$

En d'autres termes le vecteur b est calculé à partir d'une opération très simple.

On considère le vecteur $\mathcal{A}$ des coefficients $a_1, ..., a_i, ..., a_n$, on enlève la k-ième composante a_k ce qui nous donne le vecteur $\mathcal{A}^{[k]}$:

$$\mathcal{A}^{[k]} = \{a_1, ..., a_{k-1}, a_{k+1}, ..., a_n\}'.$$

On prend ensuite le k-ième vecteur colonne $V^{(k)}$ de la matrice V^{-1}, on enlève la k-ième composante $v^{(k,k)}$, ce qui nous donne le vecteur $V^{[k](k)}$ à $n-1$ dimensions :

$$V^{[k](k)} = \{v^{(1,k)}, ..., v^{(k-1,k)}, v^{(k+1,k)}, ..., v^{(n,k)}\}'.$$

Le reste est trivial. On retranche de $\mathcal{A}^{[k]}$ le vecteur $V^{[k](k)}$ multiplié par le rapport $\frac{a_k}{v^{(k,k)}}$.

On peut écrire explicitement l'expression de chaque composante du vecteur :

$$b_i = \begin{cases} a_i - \frac{a_k}{v^{(k,k)}} v^{(i,k)}, & si \ \ i \prec k, \\ a_{i+1} - \frac{a_k}{v^{(k,k}} v^{(i+1,k)}, & si \ \ i \succeq k. \end{cases}$$

Nous démontrerons ce théorème, dans le cas où $k = 1$ (i.e. c'est x_1 qui est absent), le cas général s'en déduisant aisément à l'aide de matrices de permutation intervertissant les rôles des indices 1 et k. La démonstration est basée sur les formules de Frobénius relatives à l'inversion des matrices carrées en 4 blocs. Rappelons ce résultat :

Lemme 4 *Soit $R = \begin{pmatrix} r_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$, une matrice $(n \times n)$ partitionnée en 4 blocs, r_{11} étant un scalaire et R_{22} une matrice $(n-1) \times (n-1)$. On partitionne sa matrice inverse R^{-1} exactement de la même façon :*

$$R^{-1} = \begin{pmatrix} r^{(1,1)} & R^{(1,2)} \\ R^{(2,1)} & R^{(2,2)} \end{pmatrix}.$$

On a alors les relation suivantes :

$r^{(1,1)} = 1/[r_{11} - R_{12}R_{22}^{-1}R_{21}]$;

$R^{(1,2)} = -r^{(1,1)}R_{12}R_{22}^{-1}$;

$R^{(2,1)} = -r^{(1,1)}R_{22}^{-1}R_{21}$;

$R^{(2,2)} = R_{22}^{-1} + r^{(1,1)}R_{22}^{-1}R_{21}R_{12}R_{22}^{-1} = R_{22}^{-1} + R^{(2,1)}R^{(1,2)}/r^{(1,1)}$.

Posons

$$V = \begin{pmatrix} v_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix} \quad , \quad V_{Xy} = \begin{pmatrix} v_{1y} \\ V_{2y} \end{pmatrix},$$

Le vecteur b des coefficients de régression associé au vecteur $X^{[1]}$ est défini par l'expression :

$$b = V_{X^{[1]}X^{[1]}}^{-1} V_{X^{[1]}y} = V_{22}^{-1} V_{2y}.$$

Par définition :

$$a = V^{-1} V_{Xy}.$$

Compte-tenu du Lemme, on a :

$a_1 = v^{(1,1)}v_{1y} + V^{12}V_{2y}$

$a^{[1]} = V^{21}v_{1y} + V^{22}v_{2y} = V^{21}v_{1y} + (V_{22}^{-1} + V^{21}V^{12}/v^{(1,1)})v_{2y}$

$= b + V^{21}(v_{1y} + v^{12}V_{2y}/v^{(1,1)}) = b + \frac{v^{21}}{v^{(1,1)}}a_1$.

Soit,

$$b = a^{[1]} - V^{[1](1)}a_1/v^{(1,1)}$$

puisque V^{21} est égal avec nos conventions à $V^{[1](1)}$.

Le théorème est donc démontré pour $k = 1$. On en déduit immédiatement comme on l'a déjà dit le cas général $k \neq 1$, par permutation des indices 1 et k.

Nous pouvons constater également, que le lemme peut non seulement servir à démontrer le théorème, mais encore à réaliser la procédure dans le cas où il y a, non pas une, mais plusieurs données manquantes.

En effet supposons qu'il y ait l données manquantes $x_{i_1}, ..., x_{i_l}$. Nous appliquons le théorème une première fois et nous obtenons le vecteur des coefficients de régression à $n - 1$ dimensions. Maintenant pour pouvoir appliquer le théorème une deuxième fois il nous faudrait connaitre la matrice inverse de covariance du vecteur $X^{[i_1]}$, c'est-à-dire $V_{X^{[i_1]}X^{[i_1]}}^{-1}$, ce que le lemme nous permet d'obtenir sans faire d'inversion, après permutation des indices 1 et i_1.

Nous pouvons alors recommencer la procédure pour la deuxième donnée manquante puisque nous disposons des deux informations nécessaires : le vecteur des coefficients de régression à $(n - 1)$ dimensions et la matrice inverse de covariance correspondante. Nous procédons ainsi de proche en proche jusqu'à épuisement de toutes les données manquantes.

Remarque : dans le cas de la discrimination linéaire à 2 groupes, on obtient de la même façon des formules de recalcul analogues à celles données dans le cas de la régression.

7. Conclusion

Le traitement des fichiers incomplets place le statisticien devant certaines difficultés liées au fait que les formules classiques ne sont plus applicables. La non prise en compte de la spécificité du traitement des fichiers incomplets peut entrainer l'apparition de résultats incohérents. L'exemple que nous avons présenté illustre bien ce fait en conduisant à des valeurs du coefficient de corrélation supérieures à 1 ou inférieures à - 1.

Dans cet article des réponses sont apportées à quelques problèmes d'une grande utilité pratique :

- 1) l'estimation non biaisée du coefficient de covariance (6) :

$$\tilde{v}_{ij} = \frac{T_i T_j}{T_{ij}[T_i T_j - T_i - T_j + T_{ij}]} \sum_{t \in T_{ij}} [x_i(t) - \bar{x}_i][x_j(t) - \bar{x}_j],$$

- 2) la variance de cette estimation (8) :

$$\mathcal{D}(\tilde{v}_{ij}) = \frac{2v_{ij}}{T_{ij}(\theta_{ij} + T_i T_j)^2}[T_i^2 T_j^2 + 2T_i T_j \theta_{ij} + T_{ij}\theta_{ij}^2]$$

  avec $\theta_{ij} = T_{ij} - (T_i + T_j)$.

- 3) les coefficients non biaisées d'autocovariance et d'autocorrélation calculés à partir d'un fichier complet (9 et 10) :

$$v_Z(\tau) = \frac{T^2}{(T-\tau)[T^2 - T - \tau]} \sum_{t=1}^{T_\tau} [Z(t) - \bar{Z}][Z(t+\tau) - \bar{Z}] \ ,$$

$$r_Z(\tau) = \frac{T^2(T-1)}{(T-\tau)(T^2 - T - \tau)} \frac{\sum_{t=1}^{T-\tau}[Z(t) - \bar{Z}][Z(t+\tau) - \bar{Z}]}{\sum_{t=1}^{T}[Z(t) - \bar{Z}]^2} \ .$$

- 4) la détermination d'une vraie matrice de covariance "la plus proche" au sens d'une métrique quadratique naturelle du tableau $\tilde{V}$ formé par la juxtaposition des coefficients de covariance calculés séparément pour chaque couple de variables :

$$\hat{V} = arg \quad \underset{\forall V \in def(\succeq 0)}{min} \ Tr[\tilde{V} - V]^2 = C\Lambda^+ C'.$$

- 5) le recalcul automatique des coefficients de régression ou de discrimination quand l'un ou plusieurs prédicteurs sont manquants (13) :

$$b = \mathcal{A}^{[k]} - \frac{a_k}{v^{(k,k)}} V^{[k](k)}.$$

Les solutions apportées permettent le traitement correct des fichiers perturbés par les données manquantes.

8. Références bibliographiques

[1] AIVAZIAN, MECHALKINE, ENIOUROV : Statistique Appliquée. MIR, MOSCOU, 1984.

[2] ANDERSON : An introduction to multivariate statistical analysis. WILEY, N.Y., 1958.

[3] BALESTRA : La dérivation matricielle. Collection de l'Inst. de Math. Econ. de l'Un. de DIJON, n^o12, DIJON, 1975.

[4] BELOOUSOV, GANDIN, MACHKOVITCH : Le traitement de l'information météorologique opérationnelle à l'aide des ordinateurs. HYDROMETEOIZDAT, LENINGRAD, 1968 (en russe).

[5] DER MEGREDITCHIAN : Méthodes statistiques d'analyse et d'interpolation spatiales des champs météorologiques. Note OMM (à paraître).

[6] DER MEGREDITCHIAN : Quelques aspects de la sélection des prédicteurs en analyse discriminante. La Météorologie, n^o6, 1979.

[7] DER MEGREDITCHIAN : La prévision statistique des phénomènes météorologiques. Note EERM n^o100. PARIS, 1981.

[8] DER MEGREDITCHIAN : Une identité algébrique utile pour la statistique mathématique. La Météorologie, Série II, n^o32, Mars 1983.

[9] DER MEGREDITCHIAN : Le traitement statistique des données multidimensionnelles. E.N.M. - TOULOUSE 1981.

[10] DER MEGREDITCHIAN : Problèmes engendrés par les données manquantes dans la pratique statistique. Note EERM n^o208. PARIS, 1988.

[11] DRAPER, SMITH : Applied Regression Analysis. Wiley, N.Y., 1966.

[12] GANDIN : Objective analysis of meteorological fields. ISRAEL. TRANSL. PROGR., JERUSALEM, 1965.

[13] GANDIN, KAGAN : Méthodes statistiques d'interprétation des données météorologiques. HYDROMETEOIZDAT, LENINGRAD, 1974 (en russe).

[14] MORRISON : Multivariate statistical methods. Mc GRAW HILL, N.Y., 1967.

[15] RAO : Linear statistical inference. WILEY, N.Y., 1981.

[16] RAO : Advanced statistical méthods in biometric research. WILEY, N.Y., 1962.

[17] SEBER : Linear Regression Analysis. WILEY, N.Y., 1977.

[18] WILKS : Mathematical statistics. WILEY, N.Y., 1962.