LES CAHIERS DE L'ANALYSE DES DONNÉES

K. BEN SALEM

Une condition suffisante pour l'existence de valeurs finies des données manquantes satisfaisant au critère de la trace minima

Les cahiers de l'analyse des données, tome 18, n° 3 (1993), p. 369-372

<http://www.numdam.org/item?id=CAD_1993__18_3_369_0>

Les Cahiers de l'Analyse des Données
Vol XVIII - 1993 - n°3 - pp.369-372.

UNE CONDITION SUFFISANTE POUR L'EXISTENCE DE VALEURS FINIES DES DONNÉES MANQUANTES SATISFAISANT AU CRITÈRE DE LA TRACE MINIMA

[FIN. TRAC. MIN.]

*K. Ben SALEM**

1 Position du problème

Dans [TRAC. MANQ.] (in *CAD*, Vol.XVII, n°1, 1992), on suggère de reconstituer les données manquantes d'un tableau de correspondance, d'après le "*critère de la trace minima*". Le problème n'est complètement résolu que dans le cas d'une donnée manquante unique: il y a, alors, existence et unicité d'une valeur finie satisfaisant au critère, sous réserve d'une condition énoncée dans [TRAC. MANQ.], §4, et étudiée en détail dans [PROB. TRAC. MIN.], §2.7.

Le cas de plusieurs données manquantes a été abordé dans [TRAC. VAL. NUL.] (in *CAD*, Vol.XVII, n°4, 1992), sous des hypothèses très restrictives, qui permettent, toutefois, d'apprécier les difficultés rencontrées.

On se propose ici d'énoncer une condition suffisante pour l'existence (mais non l'unicité) d'un système de valeurs finies satisfaisant au critère. En bref, la condition, précisée ci-après, généralise celle énoncée dans [TRAC. MANQ.].

Mais il faut préciser ce qu'on entend par valeurs finies: dans [TRAC. MANQ.], on prend pour objet inconnu, non le tableau k_{IJ} complété, mais la loi de probabibilité, t_{IJ}, associée à celui-ci; ainsi, le paramètre inconnu, noté x, est nécessairement compris entre 0 et 1; ce qui permet de raisonner dans un domaine borné, même si, en fait, la valeur x=1 correspond à $k(i_o, j_o) = \infty$, *infini*, qui n'est pas une solution acceptable. Ici, nous prendrons directement pour inconnues les valeurs manquantes, $\{k(i_a, j_a) \mid a \in A\}$, chacune étant supposée comprise dans la demi-droite fermée $[0, \infty]$, ce qui n'offrira pas d'obstacle à l'analyse parce que nous admettrons qu'on effectue sur chacune des inconnues une transformation homographique, la réduisant à un intervalle fini.

(*) Département des Sciences de l'Informatique; Faculté des Sciences de Tunis.

2 Hypothèses sur la structure du tableau: place des données manquantes et valeurs des données spécifiées

On distingue, dans $I \times J$, relativement au tabeau k_{IJ} considéré, d'une part, des cases (i, j) où figurent des données spécifiées; et, d'autre part, des cases, encore appelées cases vides, où la donnée manque.

On note:

$k_{IJ} = \{k(i, j) \mid i \in I, j \in J\}$: le tableau considéré, où toutes les valeurs spécifiées sont ≥ 0, puisqu'il s'agit d'un tableau de correspondance ;

$k_0(i)$, $k_0(j)$: les valeurs marginales calculées en mettant à zéro toutes les données manquantes.

$A = \{(i_a, j_a) \mid a \in A\}$: l'ensemble des cases où la donnée manque.

$I_A = \{i_a \mid a \in A\}$; $J_A = \{j_a \mid a \in A\}$: respectivement, ensemble des lignes et ensemble des colonnes comportant une case vide.

$I_+ = I - I_A$; $J_+ = J - J_A$; respectivement, ensemble des lignes et ensemble des colonnes ne comportant aucune case vide.

On suppose:

que toute case vide, (i_a, j_a), est seule sur sa ligne et seule sur sa colonne; autrement dit, que les applications $a \rightarrow i_a$ et $a \rightarrow j_a$ sont des injections de A, dans I et J respectivement ; avec inclusion stricte, i.e. $I_+ \neq \emptyset$, $J_+ \neq \emptyset$;

que, plus précisément, en notant $I_{+a} = I_+ \cup \{i_a\}$, $J_{+a} = J_+ \cup \{j_a\}$, le sous-tableau de k_{IJ} croisant I_{+a} avec J_{+a}, sous-tableau qui ne comporte pour case vide que (i_a, j_a), satisfait aux conditions étudiées dans [PROB. TRAC. MIN.], §2.7; conditions rappelées ci-dessous:

que dans toute ligne i, et dans toute colonne j, il y a, au moins un élément spécifié non nul ;

qu'est satisfaite au moins l'une des deux conditions suivantes (équivalentes aux inégalités strictes $h<H$ et $g<G$) :

il existe une colonne j_+ de J_+ comportant aux moins deux données non nulles dont l'une est $k(i_a, j_+)$;

il existe une ligne i_+ de I_+ comportant aux moins deux données non nulles dont l'une est $k(i_+, j_a)$;

3 Domaine de reconstitution et paramétrage des inconnues

Chaque donnée manquante, $k(i_a, j_a)$ est, *a priori*, regardée comme un élément de la demi-droite fermée $[0, \infty]$; reconstituer les données manquantes, c'est faire choix d'un système de valeurs qui constitue un élément de l'ensemble produit: $[0, \infty]^A$; on appellera cet ensemble, le *domaine de reconstitution*; topologiquement, de même que la demi-droite $[0, \infty]$ est isomorphe au segment (0, 1), le domaine de reconstitution est isomorphe à un hypercube fermé d'arête 1, à cardA dimensions.

Le problème est d'étudier la variation du critère sur ce domaine. On se propose de démontrer que, sous les hypothèses formulées au §2, le critère atteint effectivement son minimum en au moins un point du domaine; et que, plus précisément, le minimum ne peut être atteint qu'en un point intérieur à l'hypercube; i.e., que les valeurs attribuées aux données manquantes sont toutes, nécessairement finies et non nulles. Ainsi se trouvera généralisé le résultat d'existence établi au §4 de [TRAC. MANQ.] pour le cas d'une seule donnée manquante; mais non l'unicité, dont nous ne savons rien.

Ainsi qu'on l'a annoncé au §1, le domaine de variation sera paramétré de telle sorte que l'étude du critère devienne l'étude d'une fonction continue bornée différentiable, et même algébrique, sur le cube $(0, 1)^A$ lui-même. À cette fin, on substitue à chacune des inconnues $k(i_a, j_a)$ une variable x_a qui en est une fontion homographique et varie de 0 à 1 quand $k(i_a, j_a)$ varie de 0 à ∞.

Ceci peut être fait de plusieurs manières; dans la suite, on posera: $x_a = k(i_a, j_a) / (k(i_a, j_a)+k_0(i_a))$; i.e., on prendra pour paramètre la composante f_{ja}^{ia} du profil de la ligne i_a ; on aurait pu prendre $y_a = k(i_a, j_a) / (k(i_a, j_a)+k_0(j_a))$, composante f_{ia}^{ja} du profil de la colonne j_a . Il importe de remarquer que x_a et y_a sont fonctions homographiques l'un de l'autre; et que, sur un voisinage de l'intervalle utile (0, 1) la correspondance est biunivoque et dérivable dans les deux sens.

Le paramétrage étant fixé, il reste à montrer que le critère est une fonction algébrique des x_a, continue, bornée, différentiable, sur le cube, et même sur un voisinage ouvert de celui-ci. Il suffit de montrer que chacun des termes $Crit(i,j) = k(i,j)^2/(k(i).k(j))$ possède ces propriétés. Pour un terme $Crit(i_a, j_a)$, c'est clair: car il n'est autre que le produit $x_a.y_a$ du paramètre par une fonction homographique de celui-ci. Pour un terme (i,j) dont la valeur k(i,j) est spécifiée, chacun des facteurs (k(i,j)/k(i)) et (k(i,j)/k(j)) est, soit constant (s'il n'y a pas de donnée manquante respectivement dans la ligne i ou la colonne j), soit fonction homographique, à valeur dans (0,1), de l'un des paramètres adoptés; e.g., si $i=i_a$, $k(i,j)/k(i) = k(i_a,j)/(k_0(i_a)+k(i_a,j_a))$, avec $k(i_a,j) \leq k_0(i_a)$.

4 Réalisation du minimum du critère à l'intérieur du domaine de définition

Puisque sur le cube paramétré par les cardA variables $x_a \in (0, 1)$, le critère est une fonction continue, le minimum du critère est certainement atteint: il reste à montrer qu'il ne peut l'être qu'en un point intérieur; i.e., pour des valeurs des x_a, et donc des valeurs reconstituées $k(i_a, j_a)$, toutes comprises dans l'intervalle ouvert)0, 1(. Or des hypothèses suffisantes ont été posées, au §2, pour que la démonstration se ramène, sans nouveau calcul, à ce qui a été démontré dans le cas d'une seule donnée manquante.

Supposons que le minimum du critère (sur le cube) est réalisé pour une combinaison particulière $\{x_a \mid a \in A\}$ des valeurs des paramètres. Il faut montrer que chacun des x_a est intérieur à l'intervalle ouvert)0, 1(; ce qui équivaut à dire que chacune des données manquantes $k(i_a, j_a)$ a reçu une valeur finie strictement positive.

À cette fin, prenons une réalisation quelconque du minimum de Crit; sans supposer, *a priori*, que les valeurs attribuées aux données manquantes ne puissent être nulles ou infinies. Et considérons une case vide particulière, notée (i_a, j_a), dont nous faisons varier le contenu $z_a = k(i_a, j_a)$, tandis que celui des autres cases reste fixé. Il nous suffit de montrer que, en fonction de z_a, supposé ≥ 0, Crit atteint effectivement son minimum pour une valeur de z_a finie et non nulle.

Or, des termes Crit(i,j) dont Crit est la somme, tous sont indépendants de z_a, exceptés ceux afférents à la ligne i_a ou à la colonne j_a. Parmi ceux-ci, certains pourraient se rapporter à une colonne ou une ligne contenant une case ayant reçu la valeur ∞; (éventualité qui ne peut se réaliser, mais que nous ne pouvons encore écarter). Ces termes sont certainement nuls et l'on peut, sans modifier $Crit(z_a)$, sinon par une constante, remplacer par 0 les k(i,j) correspondants. Ceci fait, les valeurs infinies introduites dans k_{IJ} peuvent être remplacées, e.g., par 1; toujours sans altérer $Crit(z_a)$ de plus que d'une constante.

Le tableau k_{IJ} ainsi modifié contient certainement, comme sous-tableau non modifié, le tableau $I_{+a} \times J_{+a}$ dont on a postulé les propriétés au §2; et il possède *a fortiori* ces propriétés, car, e.g., s'il existe une colonne j_+ de J_+ comportant au moins deux données non nulles dont l'une est $k(i_a, j_+)$, celà reste vrai dans k_{IJ} modifié. Il en résulte que le minimum de Crit ne peut être atteint que pour z_a fini non nul. CQFD.

Nous conclurons en rappelant que des conditions d'unicité restent à trouver.