REVUE DE STATISTIQUE APPLIQUÉE

A. ANTONI
T. DHORNE

Information des tableaux individus x variables

Revue de statistique appliquée, tome 43, n° 4 (1995), p. 43-61
<http://www.numdam.org/item?id=RSA_1995__43_4_43_0>

Rev. Statistique Appliquée, 1995, **XLIII** (4), 43-61

INFORMATION DES TABLEAUX INDIVIDUS × VARIABLES

A. Antoni (1), T. Dhorne (2)

(1) Institut Universitaire de Technologie
Laboratoire d'Applications et de Méthodologies Statistique et Informatique
8 rue Montaigne - 56014 Vannes
(2) Institut National de la Recherche Agronomique
Laboratoire de Biométrie
65 rue de St Brieuc - 35042 Rennes Cedex

RÉSUMÉ

Le problème de la sélection d'un sous-ensemble de variables pertinentes dans un ensemble de variables mesurées sur plusieurs individus peut être analysé en terme de réduction de l'information. On présente un cadre formel définissant l'information de tableaux individus × variables. On montre ensuite que les approches classiques de l'analyse factorielle sont souvent inadaptées pour le problème. Certaines mesures plus adaptées sont proposées et validées sur deux jeux de données : l'un artificiel, l'autre classique de l'analyse des données.

Mots-clés : *Analyse Factorielle, Information.*

ABSTRACT

The problem of the selection of variables subsets in a set of variables measured on individuals can be analysed through reduction of information. A formal approach defining the information of two-way arrays is presented. It is shown that the classical approaches of factor analysis are seldom adapted to the problem. Some measures more convenient are proposed and worked on two data sets : one artificial and the other classical in data analysis.

Keywords : *Factor Analysis, Information*

1. Introduction

La recherche de sous-ensembles «représentatifs » d'un ensemble de variables mesurées sur une collection d'individus a donné lieu à divers travaux [Beale *et al.* 75], [Jolliffe 72], [Jolliffe 73], [McCabe 84], [Krzanowski 87]. Cependant un des problèmes posés par ces approches est celui de la définition du résumé d'information obtenu et, par là même, de l'information initiale du tableau «individus × variables »

Les mesures d'informations disponibles à l'heure actuelle ne permettent pas une recherche efficace de sous-ensembles de variables représentatives. En effet, les travaux concernant les mesures de l'information de tableaux individus × variables relèvent essentiellement de deux types d'approches :

– les approches issues des méthodes d'analyses factorielles,

– les approches concernant l'information de distributions multivariables.

dont aucune n'est bien adaptée à l'objectif recherché.

On connaît le lien formel qui existe entre l'information de Shannon et la mesure du lien en Analyse des Correspondances [Volle 84] et [Ulmo Bernier 73]. Mais cette propriété n'est pas transposable en Analyse en Composantes Principales qui est la méthode de référence pour l'étude de tableaux individus × variables.

Pour ce qui concerne le résumé de l'information, les travaux de [Braun 73] par exemple, montrent que le résumé d'information de l'Analyse en Composantes Principales n'est pas optimal et suggèrent des améliorations. De même, [Mallet 71] propose des transformations visant à améliorer les possibilités d'interprétation fournies par cette analyse.

Les travaux concernant l'information de distributions multivariables sont très nombreux (voir [Fisher 25], [Kullback 68] [Renyi 66], [Rao 73][pp 329, 331]). Cependant le contexte probabiliste dans lequel ils se placent ne correspond pas exactement à l'objectif qui est fixé ici. De même, les méthodes pour apprécier la redondance d'information lorsque les variables sont corrélées [Der Megreditchian 90] [Sneyers 71] ne sont pas totalement adaptées. Des modifications doivent être apportées pour répondre à la spécificité du problème.

L'objectif de cet article est de proposer des mesures d'information de tableaux individus × variables. Dans un premier temps, on précise formellement les propriétés que «devraient » vérifier des mesures pertinentes d'information. On montre ensuite rapidement qu'aucune des mesures classiques ne vérifie ces propriétés, puis on en étudie un certain nombre, plus judicieuses pour l'objectif fixé. Les différentes mesures sont comparées numériquement et un exemple classique de l'analyse des données est étudié.

2. Propriétés des mesures d'information des tableaux individus × variables

Dans la suite, nous considérons un tableau de données individus × variables représenté par la matrice :

$$X = (x_{ij}).$$

L'analyse en Composantes Principales conduit à l'étude de la matrice de corrélation des variables initiales :

$$T'T,$$

où la matrice T est la matrice des variables initiales centrées et réduites.

La recherche d'une représentation factorielle du tableau X consiste en la sélection d'un sous-espace de représentation optimal au sens du critère de la trace, c'est-à-dire d'une famille de vecteurs $\{u_i, i = 1, \ldots, r\}$ tels que $\sum_{i=1}^{r} u_i' T' T u_i$ soit maximale.

Ceci revient implicitement à considérer la trace tr $(T'T)$ comme une mesure de l'information du tableau de données et à rechercher le sous-espace de représentation qui maximise l'information retenue.

Bien que cette remarque semble évidente, il importe d'insister sur le fait que l'on considère souvent le critère de la trace comme un simple critère d'ajustement, mais il correspond à notre sens aussi (et peut-être surtout) à une formulation de la «valeur informative » du tableau de données. Cet aspect intervient dans l'étude conjointe des variables et dans la prise en compte de leurs relations. La différence entre l'Analyse en Composantes Principales et l'Analyse Factorielle Discriminante est, de ce point de vue, tout à fait intéressante. Dans un cas, on mesure l'information relativement à une idée d'orthogonalité entre variables, dans l'autre on la mesure relativement à la variabilité résiduelle (intra).

Il importe donc, à travers le résumé d'information opéré, de se poser le problème de la définition d'une mesure d'information et de ses propriétés.

La matrice T est la juxtaposition de p vecteurs colonnes :

$$T = [T_1 \mid T_2 \mid \cdots \mid T_i \mid \cdots \mid T_p],$$

où $\mid$ est l'opérateur de la juxtaposition en colonne.

Soit I une application de $(R^n)^p$ dans R, mesurant l'information d'un tableau $T_{(n,p)} : I(T)$. Il paraît souhaitable que l'information vérifie les postulats suivants :

– L'information ne dépend pas de l'ordre des variables.

$$I(T) = I(P(T)), \tag{1}$$

où $P(T)$ est un tableau permuté par colonnes de T.

– L'information d'une variable de variance nulle est nulle.

$$\forall \lambda \in R, I(\lambda 1_n) = 0. \tag{2}$$

On ne peut d'ailleurs pas réduire une telle variable.

– L'information d'une variable de variance non nulle est unitaire.

$$I(T_i) = 1, \ \forall T_i \neq \lambda 1_n. \tag{3}$$

– L'ajout d'une variable ne peut diminuer l'information ni l'augmenter de plus d'une unité :

$$I(T) + 1 \geqslant I(T \mid T_+) \geqslant I(T). \tag{4}$$

- L'ajout d'une variable colinéaire aux variables déjà prises en compte ne change pas l'information :

$$T_+ \in Im(T) \Rightarrow I(T \mid T_+) = I(T). \tag{5}$$

- L'ajout d'une variable identique à une des variables déjà prises en compte ne change pas l'information :

$$\exists i \in [1, \ldots, p], T_+ = T_i \Rightarrow I(T \mid T_+) = I(T). \tag{6}$$

  Cette propriété est une forme affaiblie de (5), elle correspond à une propriété minimale qui doit être vérifiée par une mesure «sérieuse » de l'information, elle est comme on le verra dans la suite plus facile à atteindre que (5).
- L'ajout d'une variable orthogonale aux variables déjà connues augmente l'information unitairement.

$$T_+ \in Im(T)^{\perp} \Leftrightarrow I(T \mid T_+) = I(T) + 1. \tag{7}$$

Toutes les propriétés sont invariantes par homothétie, $I(T_i) = I(\lambda T_i)$, ce qui explique que l'on travaille sur des variables réduites sans perte de généralité.

3. Mesures algébriques de l'information

3.1. Le critère de la trace

Nous avons rappelé plus haut que l'Analyse en Composantes Principales correspondait à une mesure de l'information :

$$I(T) = \mathrm{tr}(T'T) = p, \tag{8}$$

où p est le nombre de variables.

Il est aisé de voir que cette mesure vérifie les propriétés (1), (2) et (3). Elle vérifie aussi (4) et (7) mais de façon triviale puisqu'elle ne tient pas compte de la nature de la variable rajoutée. En conséquence, elle ne peut vérifier la propriété (6).

Ce point est important à souligner car c'est à notre sens l'inconvénient majeur du critère de la trace. Le cas extrême est celui où l'on juxtapose p fois la même variable, ce qui se traduit par l'augmentation injustifiée de l'information.

3.2. Une mesure dérivée

Il n'y a guère à l'heure actuelle d'autre critère utilisé dans la recherche d'axes d'inertie que celui de la trace. Il est possible de proposer une mesure plus proche des

axiomes posés :

$$I(T) = \sum_{i=1}^{p} \inf(1, \lambda_i), \tag{9}$$

où les λ_i sont les valeurs propres de $T'T$.

Cette normalisation est d'ailleurs utilisée pour équilibrer l'inertie de sous-nuages en ACM [Escofier Pagès 89][p 134]. De manière identique à celle de l'Analyse en Composantes Principales, cette mesure d'information où les facteurs sont en quelque sorte normés, vérifie les propriétés (1),(2) et (3). La propriété (4) est une conséquence du théorème de séparation de Sturm [Rao 73][pp 62, 64]. Elle vérifie la propriété (7) de façon non triviale. En effet, l'adjonction d'une variable non orthogonale aux précédentes ne peut augmenter l'information unitairement. Cependant la propriété (5) n'est vérifiée que dans certains cas, en particulier quand la variable T_+ appartient à l'espace engendré par les vecteurs propres de $T'T$ associés aux valeurs propres supérieures à 1.

Cette mesure plus cohérente avec notre objectif présente le double avantage d'être invariante et facile à calculer.

3.3. Le critère du déterminant

Dans beaucoup de problèmes de recherche d'optimum sur des fonctions matricielles on utilise les critères de la trace et du déterminant. En analyse factorielle, le critère du déterminant est très peu utilisé bien qu'il ait des propriétés d'optimalité connues [Okamoto 69]. Le déterminant est cependant directement lié à l'information de Fisher pour les distributions multinormales [Fisher 25].

Soit le tableau T et le déterminant associé $\det(T'T)$, qui induisent sur T une mesure d'information :

$$I(T) = \det(T'T). \tag{10}$$

Cette mesure vérifie la propriété (1) ainsi que les propriétés (2) et (3) de façon évidente. L'originalité de cette mesure concerne surtout les propriétés (4), (5) et (7).

La propriété (4) ne peut être vérifiée. Le déterminant d'une matrice de corrélation est une fonction strictement décroissante du nombre de variables, en dehors de l'orthogonalité. En effet le déterminant associé du tableau $(T|T_+)$ se décompose selon la formule suivante [Rao 73][pp 32, 33] :

$$\det([T|T_+]'[T|T_+]) = \det(T'T)(1 - Q).$$

où $Q = ((T_+)'T)(T'T)^{-1}(T'T_+)$.

Q étant un scalaire positif, $\det(1 - Q) = 1 - Q$. Or comme $\det(T'T)$ et $\det([T|T_+]'[T|T_+])$ sont positifs (les matrices associées étant définies positives), on a :

$$1 - Q \geqslant 0.$$

et donc :

$$0 \leqslant (1 - Q) \leqslant 1,$$

ce qui démontre le résultat.

De plus, l'adjonction d'une variable colinéaire aux précédentes entraîne l'annulation du déterminant.

Le critère du déterminant peut donc sembler inintéressant mais il possède en fait des propriétés complémentaires à celles de la trace.

Nous avons vu plus haut que l'information de la trace ne tient nullement compte de la colinéarité, en mettant toutes les variables sur un pied d'égalité. L'information du déterminant, quant à elle, privilégie l'orthogonalité, seul critère ne dégradant pas l'information.

Ceci est dû au fait que le déterminant est le pendant pour la multiplication de la trace pour l'addition. Notre information étant, par exemple en ce qui concerne la propriété (4), implicitement additive, le déterminant n'est pas strictement adapté. Les réflexions précédentes permettent cependant de penser qu'il est judicieux d'utiliser de manière conjuguée les propriétés de la trace et du déterminant pour élaborer une mesure pertinente de l'information.

3.4. L'information des sous-déterminants

La trace et le déterminant sont les bornes d'une famille indicée d'opérateurs définis sur une matrice carrée. Cette famille est celle des traces des produits extérieurs de l'endomorphisme associé $T'T$, ou encore celle des sommes de mineurs ou sous-déterminants.

On appelle sous-déterminant d'ordre $i, i \in [1, \ldots, p]$ d'une matrice carrée M d'ordre p, le déterminant d'une sous-matrice $M^{(i)}$ d'ordre i de M. Il est immédiat que le sous-déterminant d'ordre p est le déterminant de M et que la somme des sous-déterminants d'ordre 1 est la trace de M.

Soit $\mathcal{C}(i,p)$ l'ensemble des C_i^p combinaisons de i éléments pris parmi p, c une de ces combinaisons, T^c le tableau constitué à partir de T en ne retenant que les colonnes de c, $\det({T^c}'T^c)$ est un sous-déterminant d'ordre i de $T'T$ et $\sum_{c \in \mathcal{C}(i,p)} \det({T^c}'T^c)$ la somme de tous les déterminants d'ordre i.

Nous nous intéressons dans la suite à des combinaisons linéaires de sous-déterminants de la matrice $T'T$. Ces mesures d'information vérifient comme le déterminant les propriétés (1) et (2).

3.4.1. L'information totale des sous-déterminants

La combinaison linéaire suivante est donc d'intérêt :

$$I(T) = \sum_{i=1}^{p} k_i^p \sum_{c \in \mathcal{C}(i,p)} \det({T^c}'T^c), \tag{11}$$

où p est l'ordre de $T'T$, k_i^p une famille de constantes à définir en fonction des propriétés à respecter.

Le respect de la propriété (3) entraîne que $k_1^1 = 1$.

En cas d'ajout d'une variable, l'information peut se décomposer en (voir annexe A) :

– dans le cas où T_+ est une des variables de T :

$$I(T \mid T_+) = 2\sum_{i=1}^{p} k_i^{p+1} \sum_{c \in \mathcal{C}(i,p)} \det(T^{c\prime} T^c) - \sum_{i=1}^{p-1} k_i^{p+1} \sum_{c \in \mathcal{C}(i,p-1)} \det(T^{c\prime} T^c). \tag{12}$$

Une partie de l'information est alors dupliquée.

– dans le cas où T_+ est orthogonal à $Im(T)$ (propriété 7) :

$$I(T \mid T_+) = \sum_{i=1}^{p} (k_i^{p+1} + k_{i+1}^{p+1}) \sum_{c \in \mathcal{C}(i,p)} \det(T^{c\prime} T^c) + k_1^{p+1}. \tag{13}$$

Pour se rapprocher des propriétés (6) et (7), l'une des solutions possibles consiste à retrouver l'information antérieure dans les seconds termes des égalités (12) et (13) en posant :

$$k_i^p = 2k_i^{p+1},$$
$$k_i^p = k_i^{p+1} + k_{i+1}^{p+1}.$$

Alors k_1^{p+1} représente l'augmentation de l'information en (13).

Il s'ensuit que :

$$k_i^{p+1} = k_{i+1}^{p+1},$$

et que par conséquent k_i^p ne dépend pas de i. Les contraintes se résument à : $2k^{p+1} = k^p$ et avec la contrainte $k^1 = 1$:

$$k^p = \frac{1}{2^{p-1}}.$$

Une des conséquences est que, en cas d'orthogonalité de la variable rajoutée, l'information n'augmente pas de 1 mais de $\frac{1}{2^p}$.

Lorsque les p variables sont orthogonales, l'information totale n'est pas égale à p mais à :

$$\sum_{i=0}^{p-1} \frac{1}{2^i} = 2 - \frac{1}{2^{p-1}}.$$

Pour retrouver p, il suffit de composer avec la fonction réciproque de f qui à p associe $y = f(p) = 2 - \frac{1}{2^{p-1}}$, soit :

$$f^{-1}(y) = -\frac{\ln(1 - \frac{y}{2})}{\ln 2}.$$

On choisit finalement :

$$I(T) = -\frac{\ln(1 - \frac{1}{2^p} \sum_{i=1}^{p} \sum_{c \in \mathcal{C}(i,p)} \det(T^{c\prime} T^c))}{\ln 2}. \tag{14}$$

L'information ainsi définie ne vérifie cependant pas la propriété (4).

D'autres normalisations peuvent être proposées, par exemple :

$$k_i^p = \frac{1}{iC_i^p},$$

où $C_i^p = \frac{p!}{i!(p-i)!}$ est le nombre de combinaisons de i éléments pris parmi p. Mais il s'ensuit les mêmes problèmes en cas d'orthogonalité de la variable rajoutée que précédemment (l'information augmente de $\frac{1}{p+1}$).

3.4.2. L'information du polynôme caractéristique

Les mesures précédentes sont très liées au polynôme caractéristique de l'endomorphisme associé à la matrice $T'T$:

$$\delta_{T'T}(\alpha) = \det(T'T - \alpha I_p). \tag{15}$$

Ce polynôme peut s'écrire :

$$\begin{aligned} \delta_{T'T}(\alpha) = (-1)^p \alpha^p + (-1)^{p-1} f_1(T'T) \alpha^{p-1} + \cdots \\ + (-1)^{p-k} f_k(T'T) \alpha^{p-k} + \cdots + f_p(T'T), \end{aligned} \tag{16}$$

avec $f_k(T'T) = \mathrm{Tr} A^k(T'T)$ où $A^i(T'T)$ désigne la puissance extérieure $i^{ième}$ de l'endomorphisme $T'T$ et :

$$\begin{aligned} f_1(T'T) &= \mathrm{tr}(T'T), \\ f_p(T'T) &= \det(T'T). \end{aligned}$$

[Chambadal Ovaert 68][pp 283].

La valeur de $\delta_{T'T}(\alpha)$ dépend simplement des valeurs propres λ_i, $i = 1, \ldots, p$ de la matrice $T'T$. En effet les valeurs propres de $T'T - \alpha I_p$ sont égales à :

$$\lambda_i - \alpha, \ i = 1, \ldots, p,$$

et donc, on posera :

$$I(T) = \delta_{T'T}(\alpha) = \prod_{i=1}^{p}(\lambda_i - \alpha). \tag{17}$$

Cette famille de mesures vérifie la propriété (1). Lorsque l'on ajoute une variable orthogonale aux précédentes, apparaît une nouvelle valeur propre égale à 1 dans la matrice $T'T$. Le polynôme caractéristique est donc multiplié par $1 - \alpha$.

Pour vérifier une contrainte additive, on peut choisir une transformation de la mesure de la forme suivante :

$$I(T) = k\ln(\delta_{T'T}(\alpha)), \tag{18}$$

où α est tel que $\delta_{T'T}(\alpha) > 0$. Dans ce cas, pour vérifier la propriété (7), k et α sont liés par la relation :

$$k\ln(\delta_{(T|T_+)'(T|T_+)}) = k\ln(\delta_{T'T}) + 1,$$

soit :

$$k\ln(1 - \alpha) = 1.$$

On prendra par exemple, $k = 1$ et $\alpha = 1 - e$ ou $k = \dfrac{1}{\ln 2}$ et $\alpha = -1$. Ces deux formes vérifient les propriétés (3) et (4), la seconde a l'avantage de vérifier la propriété (2). Cette dernière forme est très proche de (14) du paragraphe précédent. Les propriétés (5) ou (6) ne peuvent être vérifiées car les valeurs propres étant modifiées, le produit $\prod_{i=1}^{p}(\lambda_i - \alpha)$ est modifié sauf pour une valeur de α tendant vers $+\infty$, ce qui n'a pas d'intérêt ici.

3.4.3. L'information de la somme des maximums des sous-déterminants

Pour pallier les inconvénients des mesures d'information précédentes, on peut utiliser la nouvelle mesure :

$$I(T) = \sum_{i=1}^{p} \max_{c \in \mathcal{C}(i,p)} \det(T^{c\prime}T^c), \tag{19}$$

avec les mêmes notations que précédemment. Comme pour la mesure définie par (11) ou (14), les propriétés (1), (2) et (3) sont immédiatement vérifiées. La propriété (4) est vérifiée car :

$$\max_{c \in \mathcal{C}(i,p)} \det(T^{c\prime}T^c) \leqslant \max_{c \in \mathcal{C}(i,p+1)} \det([T \mid T_+]^{c\prime}[T \mid T_+]^c).$$

L'égalité est vérifiée, en particulier, lorsque T_+ est égale à une des T_i. En effet, l'ajout de la variable T_+ entraîne la nullité de tous les déterminants des sous-tableaux

contenant T_i et T_+. Par ailleurs, les autres sous-déterminants non nuls sont les mêmes que ceux construits à partir de T et donc leur maximum est inchangé. Il s'ensuit que la somme des maximums des sous-déterminants est inchangée et la propriété (6) est ainsi vérifiée.

En ce qui concerne la propriété (7), le déterminant d'une matrice de corrélation $T'T$ est une fonction décroissante de l'ordre de T et donc :

$$\max_{c\in\mathcal{C}(i,p)} \det({T^c}'T^c) \geqslant \max_{c\in\mathcal{C}(i+1,p+1)} \det([T \mid T_+]^{c\prime}[T \mid T_+]^c),$$

l'égalité n'étant vérifiée que si T_+ est orthogonal à $Im(T)$. Dans ce cas, on a donc :

$$\begin{aligned} I(T \mid T_+) &= \sum_{i=1}^{p+1} \max_{c\in\mathcal{C}(i,p+1)} \det([T \mid T_+]^{c\prime}[T \mid T_+]^c), \\ &= \max_{c\in\mathcal{C}(1,p+1)} \det([T \mid T_+]^{c\prime}[T \mid T_+]^c) + \sum_{i=2}^{p+1} \max_{c\in\mathcal{C}(i-1,p)} \det({T^c}'T^c), \\ &= 1 + \sum_{i=1}^{p} \max_{c\in\mathcal{C}(i,p)} \det({T^c}'T^c). \end{aligned}$$

et la propriété (7) est donc vérifiée.

La seule mesure proposée jusqu'ici et vérifiant l'ensemble des propriétés souhaitées est celle de la somme des maximums des sous-déterminants (la propriété (5) ne l'est pas mais sa forme affaiblie (6) l'est). Dans la suite, nous allons considérer 2 mesures proches possédant les mêmes propriétés mais plus faciles à calculer.

3.5. Mesures ascendantes ou descendantes optimales

La somme des maximums des sous-déterminants est telle que les maximums sont définis de façon absolue. Cependant aucune des propriétés ne l'impose, on peut donc étudier des maximums contraints par exemple en imposant qu'un sous-ensemble d'ordre i contienne un sous-ensemble d'ordre $j < i$. La mesure induite a les mêmes propriétés que la somme des maximums des sous-déterminants pour des raisons évidentes et correspond dans le cas d'une inclusion ascendante à la somme des informations conditionnelles liées au déterminant. En effet, soit le déterminant du tableau complété $T \mid T_+$, on a :

$$\begin{aligned} \det([T \mid T_+]'[T \mid T_+]) &= \det(T'T) \times \mid T'_+T_+ - T'_+T(T'T)^{-1}T'T_+ \mid, \\ &= \det(T'T) \times (1 - r^2(T_+, T)). \end{aligned}$$

où $r^2(Y, X)$ est le coefficient de détermination de la régression de Y sur les variables X. Il s'ensuit que la maximisation de $\det([T \mid T_+]'[T \mid T_+])$ à T fixé consiste à choisir

T_+ la moins expliquée par T, de proche en proche on a donc :

$$I(T) = 1 + (1 - r^2(T_{(2)}, T_{(1)}) + (1 - r^2(T_{(3)}, T_{(1)} \mid T_{(2)})) + \cdots + (1 - r^2(T_{(p)}, T_{(1)} \mid \cdots \mid T_{(p-1)})). \tag{20}$$

où les $T_{(i)}$ sont ordonnés selon un ordre intrinsèque (et donc vérifiant la propriété (1)), ici l'ordre des déterminants décroissants.

Les résultats sont identiques pour un emboîtement descendant.

Ces mesures sont donc une «version additive » du déterminant, elles ont l'avantage certain de ne pas poser de problèmes algorithmiques et de permettre un calcul beaucoup plus rapide. On pourra être amené à les utiliser pour cette raison.

3.5.1. Mesure invariante induite par l'information des maximums des sous-déterminants

Un certain nombre des mesures d'information précédentes sont invariantes par changement de base : la trace, le déterminant, le polynôme caractéristique. En revanche, l'information des maximums des sous-déterminants n'est pas invariante par construction. Il peut cependant être intéressant d'étudier la mesure invariante induite afin de pouvoir l'utiliser dans le contexte classique des analyses factorielles.

Considérons la matrice $T'T$ dont les valeurs propres sont $\lambda_i,\ i = 1, \ldots, p$, nous appelerons mesure invariante induite par la mesure définie sur un ensemble de variables, une mesure construite identiquement mais sur l'ensemble des combinaisons linéaires de ces variables.

Dans le cas de l'information des maximums des sous-déterminants, on a :

$$I(T) = \sum_{i=1}^{p} \max_{c \in \mathcal{C}(i,p)} \det({T^c}'T^c).$$

Le maximum des sous-déterminants $\det({T^c}'T^c)$ étant égal au produit des i plus grandes valeurs propres, nous définirons la mesure invariante induite de la façon suivante :

$$I(T) = \sum_{i=1}^{p} \prod_{j=1}^{i} \lambda_{(j)}, \tag{21}$$

où les $\lambda_{(j)}$ sont les valeurs propres classées par ordre décroissant.

3.6. Mesures invariantes tronquées

Les mesures invariantes précédemment étudiées, fonctions simples des valeurs propres $\lambda_i,\ i = 1, \ldots, n$ de la matrice $T'T$, sont liées aux combinaisons linéaires plus qu'aux variables initiales. Le fait de travailler sur des variables initiales normées est en effet difficile à transposer au cas des vecteurs propres puisque la variance d'une

combinaison linéaire normée (vecteur propre de $T'T$) est la valeur propre associée. Lorsqu'une valeur propre est supérieure à 1, on choisira de ne pas en tenir compte et donc de remplacer λ_i par $inf(1, \lambda_i)$ dans les formules précédemment définies. Ces mesures induites sont appelées mesures tronquées. Elles sont naturellement invariantes. Une mesure de ce type a été évoquée précédemment, il s'agit de la trace tronquée (9). Nous considérons aussi la mesure invariante des maximums tronquée ainsi que le polynôme caractéristique tronqué et ses mesures dérivées.

3.7. Nombre équivalent

Le nombre équivalent été introduit [Bartels 43] dans le même objectif de la prise en compte de la redondance d'information engendrée par des mesures corrélées notamment en météorologie. L'approche proposée est assez proche de celle développée ici, le nombre équivalent d'observations étant le nombre d'observations fictives indépendantes qui conduiraient à la même précision d'estimation ramené au nombre d'observations réelles, corrélées.

En se limitant aux moments d'ordre un et deux, on peut considérer les nombres équivalents n_s [Sneyers 71] :

$$n_s = \frac{\mathrm{tr}^2(V)}{\mathrm{tr}(VJ)}, \tag{22}$$

et n_m [Der Megreditchian 90] :

$$n_m = \frac{\mathrm{tr}^2(V)}{\mathrm{tr}(V^2)}, \tag{23}$$

où V est la matrice de variance covariance des observations et J la matrice des 1.

Pour prolonger le parallèle, il conviendrait de remplacer la matrice de variance-covariance par la matrice de corrélation C.

Il est facile de voir que cette information vérifie les propriétés (1), (3). Elle ne vérifie pas, en revanche, la propriété (4) ce qui peut facilement se voir sur l'exemple de 3 variables de variance x_1, x_2, x_3 telles que :

$$x_2 \perp x_1, x_3 = x_2,$$

on a alors :

$$n_s(x_1 \mid x_2) = n_m(x_1 \mid x_2) = 2,$$
$$n_s(x_1 \mid x_2 \mid x_3) = n_m(x_1 \mid x_2 \mid x_3) = \frac{3^2}{2^2 + 1^2} = \frac{9}{5}.$$

Ceci montre par ailleurs que la propriété (5) n'est pas non plus vérifiée.

Enfin, la propriété (7) n'est vérifiée que lorsque toutes les variables déjà présentes sont orthogonales. En effet, le rajout d'une variable orthogonale aux

précédentes augmente de 1 les quantités $\mathrm{tr}(C^2)$ et $\mathrm{tr}(CJ)$, d'où :

$$I(T \mid T^+) - I(T) = \frac{(p+1)^2}{f(C)+1} - \frac{p^2}{f(C)},$$
$$= \frac{(2p+1)f(C) - p^2}{f(C)(f(C)+1)},$$

où $f(C) = \mathrm{tr}(CJ)$ ou $\mathrm{tr}(C^2)$.

Pour que cette quantité soit égale à 1, il faut et il suffit que :

$$(2p+1)f(C) - p^2 = f(C)(f(C)+1),$$

Soit : $$(p - f(C))^2 = 0,$$

c'est-à-dire que toutes les variables soient orthogonales.

4. Comparaisons des mesures d'information

4.1. Comportement à l'égard d'une colinéarité croissante

Pour étudier l'évolution des différentes mesures d'information présentées précédemment, nous utilisons un tableau pour lequel la colinéarité entre variables augmente avec le nombre de variables. La matrice choisie est la matrice réduite correspondant à la matrice initiale suivante :

$$\begin{array}{ccccc} 1 & 1 & 1 & \ldots & 1 \\ 0 & 1 & 1 & \ldots & 1 \\ 0 & 0 & 1 & \ldots & 1 \\ 0 & 0 & 0 & \ldots & 1 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & 0 & \ldots & -1 \\ 0 & 0 & -1 & \ldots & -1 \\ 0 & -1 & -1 & \ldots & -1 \\ -1 & -1 & -1 & \ldots & -1 \end{array}$$

Sur ce tableau que nous choisirons à 15 colonnes et donc à 30 lignes, on calculera successivement :

– les mesures classiques :

- trace, (8)
- déterminant, (10)
- valeur du polynôme caractéristique pour $\alpha = 1$, (17) :

$$\delta(1) = \prod_{i=1}^{p} (\lambda_i - 1),$$

– les mesures transformées :

$$- I_1(T) = -\frac{\ln\left(1 - \frac{1}{2^p}\sum_{i=1}^{p}\sum_{c\in\mathcal{C}(i,p)}\det(T^{c\prime}T^c)\right)}{\ln 2}, \quad (14)$$

$$- \sum_{i=1}^{p}\ln(\lambda_i + e - 1), \text{ (18 avec } \alpha = 1 - e, k = 1)$$

$$- \frac{1}{\ln 2}\sum_{i=1}^{p}\ln(\lambda_i + 1), \left(18 \text{ avec } \alpha = -1, k = \frac{1}{\ln 2}\right)$$

– les mesures optimales :

- somme des maximums des sous-déterminants, (19)
- somme ascendante des maximums des sous-déterminants emboîtés, (20)
- somme descendante des maximums des sous-déterminants emboîtés, (20)

– la mesure invariante des sommes des maximums des sous-déterminants, (21)

– les mesures tronquées :

- trace tronquée, (9)
- somme invariante des maximums tronquée, (§3.6)

– les nombres équivalents :

- de Sneyers, (22)
- de Der Megreditchian. (23)

Les résultats de cette comparaison de mesures sont présentés en figure 1. On remarque que les 3 sommes des maximums des sous-déterminants sont très proches (ce qui est naturel sur ces données) et que la trace tronquée et la mesure invariante des maximums tronquée ne s'en écartent pas excessivement. Ceci confère à ces mesures qui ont le grand avantage d'être invariantes des qualités certaines pour mesurer l'information dans le contexte des méthodes factorielles. Les nombres équivalents de Sneyers et de Der Megreditchian semblent aussi intéressants, en revanche, ils ont l'inconvénient déjà noté et visible sur ce graphique de ne pas être non décroissants. Les autres mesures sont trop éloignées des précédentes pour être véritablement utilisables dans la pratique.

4.2. Etude d'un exemple

Nous reprendrons l'exemple des mensurations réalisées sur différents chiens tel qu'il est présenté dans [Jambu 78] référençant [De Bonis Lebeaux 74]. Nous avons uniquement sélectionné les 30 données concernant les chiens et avons appliqué notre mesure de l'information (19) au tableau des 6 variables. La somme des maximums des sous-déterminants est de 2,427 ce qui suggère que l'information contenue est

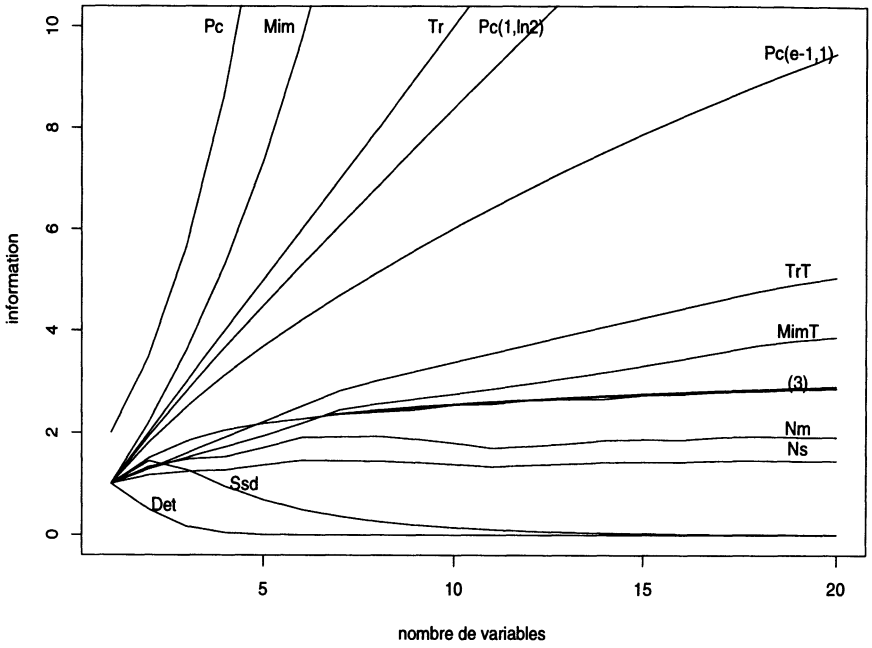

FIGURE 1 :

Comparaison des mesures d'information : en abscisse le nombre de variables et en ordonnée l'information.

Pc : polynôme caractéristique (17), Mim : Mesure invariante de la somme des maximums des sous-déterminants (21), Tr : Trace (8) , Pc(1,$\ln 2$*) et Pc(e-1,1) : polynômes caractéristiques normalisés (18), TrT : Trace tronquée (9), MimT : Mesure invariante tronquée de la somme des maximums des sous-déterminants (21), (3) : Somme des maximums des sous-déterminants (optimal, ascendant et descendant (19) , (20)) pratiquement confondus, Ssd : Somme des sous-déterminants (11), Det : Déterminant (10), Ns : Nombre équivalent de Sneyers (22), Nm : Nombre équivalent de Der Megreditchian (23).*

inférieure à celle qu'apporteraient 3 variables orthogonales. Nous avons donc choisi de ne retenir que 3 variables de ce tableau et avons étudié l'ensemble des triplés possibles. Il apparaît (voir table 1) que le sous-ensemble le plus informatif est constitué des variables 2,3 et 5 et que son information représente 96 % de l'information totale du tableau. Le sous-ensemble d'ordre 4 le plus informatif est $\{2, 3, 4, 5\}$ pour une information de 2,410. A titre comparatif, une Analyse en Composantes Principales réalisée sur ces mêmes données conduit à une inertie de 95 % sur le meilleur sous-espace de dimension 3 et de 97 % sur le meilleur sous-espace de dimension 4. La sélection des variables sur le critère de la plus forte contribution aux axes conduirait au choix du sous-ensemble $\{1, 3, 2\}$ qui dans notre approche ne fournit qu'une information de 2,037.

TABLEAU 1
Sous-ensembles de 3 variables parmi 6
et Information de la somme des maximums des sous-déterminants associée

Sous-ensemble	Information	Sous-ensemble	Information
$\{1,2,3\}$	2,037	$\{2,3,4\}$	2,312
$\{1,2,4\}$	1,456	$\{2,3,5\}$	2,332
$\{1,2,5\}$	1,421	$\{2,3,6\}$	2,286
$\{1,2,6\}$	1,509	$\{2,4,5\}$	1,528
$\{1,3,4\}$	2,257	$\{2,4,6\}$	1,601
$\{1,3,5\}$	2,219	$\{2,5,6\}$	1,589
$\{1,3,6\}$	2,223	$\{3,4,5\}$	2,158
$\{1,4,5\}$	1,480	$\{3,4,6\}$	2,089
$\{1,4,6\}$	1,515	$\{3,5,6\}$	2,128
$\{1,5,6\}$	1,497	$\{4,5,6\}$	1,386

Remerciements

Nous remercions vivement les deux lecteurs pour leur remarques pertinentes et constructives qui ont fortement contribué à l'amélioration de la version initiale.

Références

[Bartels 43] BARTELS J. 1943 *Gesetz und Zufall in der Geophysik* Naturwiss 31.

[Beale *et al.* 75] BEALE E.M.L., KENDALL M.G. and MANN D.W. 1975 *The discarding of variables in multivariate analysis.* Biometrika, 54, 3 and 4, 357-365.

[de Bonis Lebeaux 74] DE BONIS L. et LEBEAUX M.O. 1974 *A propos d'un crÉne découvert dans le quaternaire au Cantal.* Mammalia, t. 38, 717-728.

[Braun 73] BRAUN J.M. 1973 *Séries Chronologiques Multiples : Recherche d'Indicateurs.* RSA vol 21, 1, 81-106.

[Chambadal Ovaert 68] CHAMBADAL L. et OVAERT J.L. 1968 *Algèbre linéaire et algèbre tensorielle Dunod, Paris.*

[Der Megreditchian 90] DER MEGREDITCHIAN G. 1990 *Meteorological networks optimization from a statistical point of view.* Comp. Stat. and Data Anal., 9, 57-75. La Météorologie VII$^{\text{ème}}$ série, 17.

[Escofier Pagès 89] ESCOFIER B. et PAGÈS J. 1989 *Analyses factorielles simples et multiples : Objectifs, méthodes et interprétation.* Dunod, Paris.

[Fisher 25] FISHER R.A. 1925 *Theory of statistical estimation.* Proc. Camb. Phil. Soc. 700-725. [Jambu 78] JAMBU M. 1978 *Classification automatique pour l'analyse des données : méthodes et algorithmes.* Dunod, Paris

[Jolliffe 72] JOLLIFFE I.T. 1972 *Discarding Variables in a Principal Component Analysis. I : Artificial Data.* Appl. Statist., 21, 160-173.

[Jolliffe 73] JOLLIFFE I.T. 1973 *Discarding Variables in a Principal Component Analysis. II : Real Data.* Appl. Statist., 22, 21-31.

[Krzanowski 87] KRZANOWSKI W.J. 1987 *Selection of Variables to preserve Multivariate Data Structure, using Principal Components.* Appl. Statist., 1, 22-33.

[Kullback 68] KULLBACK S. 1968 *Information Theory and Statistics.* Dover Publications, New York.

[McCabe 84] MCCABE G.P. 1984 *Principal Variables.* Technometrics, 26, 1, 137-144.

[Mallet 71] MALLET J.L. 1971 *Contributions à l'Etude de Facteurs non orthogonaux en Analyse Factorielle.* RSA vol 19, 1, 57-76.

[Okamoto 69] OKAMOTO M. 1969 *Optimality of Principal Components.*, in Multivariate Analysis II, ed. P. R. Krishnaiah, Academic Press, New York, 673-685.

[Rao 73] RAO C.R. 1973 *Linear Statistical Inference and its Applications.* 2nd Edition John Wiley and sons, New York, London, Sydney, Toronto.

[Renyi 66] RENYI A. 1966 *Calcul des Probabilités avec un appendice sur la théorie de l'information.* Dunod, Paris.

[Sneyers 71] SNEYERS 1971 *Sur l'estimation du nombre équivalent de répétitions* RSA vol 19, 2, 35-47.

[Ulmo Bernier 73] ULMO J. et BERNIER J. 1973 *Eléments de décision statistique.* Presses Universitaires de France, Paris.

[Volle 84] VOLLE M. 1984 *Analyse des Données.* Economica, Paris.

A. Information des sous-déterminants

A.1. Décomposition de l'information

L'information pour p variables est :

$$I(T) = \sum_{i=1}^{p} k_i^p \sum_{c \in \mathcal{C}(i,p)} \det(T^{c\prime} T^c).$$

Etudions chacun des termes :

$$\sum_{c \in \mathcal{C}(i,p)} \det(T^{c\prime} T^c) = S(i,p), i \leqslant p,$$

et l'on posera pour assurer la cohérence :

$$S(p,q) = 0,\ \forall p > q. \tag{24}$$

Parmi les p colonnes de T, on en isole une quelconque T_j et on décompose $S(i,p)$ en :

$$S(i,p) = S(i,p-1) + S(i-1,p-1) \mid T_j, \tag{25}$$

avec :
$$S(i-1,p) \mid T_j = \sum_{c \in \mathcal{C}(i-1,p)} \det([T^c \mid T_j]'[T^c \mid T_j]),$$

qui est une somme de déterminants d'ordre $p+1$. En fait, le premier terme prend i colonnes parmi les $(p-1)$ autres que T_j et le second terme prend $(i-1)$ colonnes plus T_j.

On a de plus :
$$S(p,p) = S(p-1,p-1) \mid T_j.$$

L'ajout d'une variable aux p anciennes variables peut alors être étudiée en appliquant (25) pour $S(i,p+1)$ et la variable T_+ :

$$S(i,p+1) = S(i,p) + S(i-1,p) \mid T_+,$$

avec :
$$I(T \mid T_+) = \sum_{i=1}^{p+1} k_i^{p+1} S(i,p+1),$$
$$= \sum_{i=1}^{p+1} k_i^{p+1} \sum_{c \in \mathcal{C}(i,p+1)} \det(T^{c\prime} T^c).$$

A.2. Ajout d'une variable orthogonale

Si T_+ est orthogonal à T il s'ensuit que :

$$S(i-1,p) \mid T_+ = S(i-1,p),$$

avec la convention :

$$S(0,p) = 1, \tag{26}$$

d'où :

$$S(i,p+1) = S(i,p) + S(i-1,p),$$

et :

$$S(p+1,p+1) = S(p,p).$$

Par conséquent l'information du tableau complété se décompose en :

$$\begin{aligned} I(T \mid T_+) &= \sum_{i=1}^{p+1} k_i^{p+1} S(i,p+1), \\ &= \sum_{i=1}^{p+1} k_i^{p+1} (S(i,p) + S(i-1,p)), \\ &= \sum_{i=1}^{p} k_i^{p+1} S(i,p) + \sum_{i=1}^{p+1} k_i^{p+1} S(i-1,p), \\ &= \sum_{i=1}^{p} k_i^{p+1} S(i,p) + \sum_{i=0}^{p} k_{i+1}^{p+1} S(i,p), \end{aligned}$$

soit :

$$I(T \mid T_+) = k_1^{p+1} S(0,p) + \sum_{i=1}^{p} (k_i^{p+1} + k_{i+1}^{p+1}) S(i,p). \tag{27}$$

A.3. Ajout d'une variable identique à une des précédentes

Soit T_j la colonne de T identique à T_+. La propriété (6) n'est pas strictement vérifiée, en effet en appliquant deux fois l'égalité (25) :

$$\begin{aligned} S(i,p+1) &= S(i,p) + S(i-1,p) \mid T_+, \\ &= S(i,p) + [S(i-1,p-1) + S(i-2,p-1) \mid T_j] \mid T_+, \\ &= S(i,p) + S(i-1,p-1) \mid T_+ + S(i-2,p-1) \mid T_+ \mid T_j. \end{aligned}$$

Le dernier terme s'annulle et on a :

$$S(i,p+1) = S(i,p) + S(i-1,p-1) \mid T+,$$

où les $p-1$ variables sont différentes de la variable supplémentaire. Or nous avions vu précedemment que :

$$S(i,p) = S(i,p-1) + S(i-1,p-1) \mid T_+,$$

et donc : $$S(i,p+1) = 2S(i,p) - S(i,p-1),\ i < p.$$

Par conséquent :

$$\begin{aligned} I(T \mid T_+) &= \sum_{i=1}^{p} k_i^{p+1} (2S(i,p) - S(i,p-1)), \\ &= 2\sum_{i=1}^{p} k_i^{p+1} S(i,p) - \sum_{i=1}^{p-1} k_i^{p+1} S(i,p-1). \end{aligned} \tag{28}$$