

REVUE DE STATISTIQUE APPLIQUÉE

J. B. DENIS

Ajustements de modèles linéaires et bilinéaires sous contraintes linéaires avec données manquantes

Revue de statistique appliquée, tome 39, n° 2 (1991), p. 5-24

http://www.numdam.org/item?id=RSA_1991__39_2_5_0

© Société française de statistique, 1991, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (<http://www.sfds.asso.fr/publicat/rsa.htm>) implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/conditions>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques
<http://www.numdam.org/>

AJUSTEMENTS DE MODÈLES LINÉAIRES ET BILINÉAIRES SOUS CONTRAINTES LINÉAIRES AVEC DONNÉES MANQUANTES

J.B. DENIS

Laboratoire de Biométrie

I.N.R.A.

F78026 Versailles Cedex

RÉSUMÉ

Une modélisation linéaire et bilinéaire d'une matrice est proposée. Cette modélisation comprend des contraintes linéaires et recouvre un grand nombre de cas d'approximations pratiquées en Analyse de Données et de modèles utilisés en Analyse de la Variance à deux facteurs. Les estimations des moindres carrés des paramètres sont données lorsque la matrice est complète. Un algorithme d'estimation pour un schéma quelconque de données manquantes est proposé. Les difficultés de son emploi sont révélées, en particulier est abordé le problème de l'identifiabilité des paramètres. Enfin les résultats d'une application numérique sur des données classiques sont présentés.

Mots-clés : *covariable, approximation, moindres-carrés, données manquantes, estimabilité, analyse en composantes principales, interaction, modèle bilinéaire.*

SUMMARY

A general model for a data matrix, based on linear and bilinear functions of the parameters, is proposed. This model includes linear restrictions on the parameters. A lot of Data Analysis approximations and interaction models in two-way analysis of variance are particular cases of this model. Explicit Least-Squares estimates are given when the data matrix is complete. An Alternating Least Squares algorithm is used when there are missing values. Some of the difficulties arisen by the algorithm are considered, particularly the identifiability of model parameters. A numerical application is given.

Key-words : *covariate, approximation, Least-Squares, missing data, principal component analysis, interaction, bilinear model.*

1. Introduction

Les modèles qualifiés de Linéaires et Bilinéaires (sous-entendus en paramètres) présentés dans cet article recouvrent une assez grande variété de modèles utilisés dans le cadre de l'analyse de variance entre deux facteurs ou de formes

d'approximation en Analyse des Données. Car comme l'ont relevé Gower et Digby (1981) : «there is no distinct dividing line between data that should be considered as forming a multivariate data-matrix and data that form a two-way table». La très grande flexibilité de ces modèles autorise des traitements fins d'une matrice de données, aussi bien dans le cadre d'une analyse exploratoire que dans celui d'une analyse confirmatoire. Outre la formalisation générale, l'originalité de cette présentation réside dans la prise en compte du problème des données manquantes, abordé par la proposition d'une procédure itérative d'estimation des paramètres. Mais les formules exactes sont fournies lorsque la matrice de données est complète. Le point de vue adopté ici est celui de l'utilisation de ce type d'approximations ou de modèles dans le but de faciliter la mise au point de logiciels généraux et les résultats seront simplement énoncés. Dans le contexte de l'interaction entre deux facteurs, une mise en oeuvre fonctionnelle a été réalisée par G. Decoux et J.B. Denis (1990).

2. Définition des modèles linéaires et bilinéaires

2.1 Notations

2.1.a Matrice à Analyser

La matrice que l'on veut analyser est $\mathcal{X}$, elle comporte I lignes et J colonnes, son terme courant est généralement noté X_{ij} où i indique le numéro de la ligne et j le numéro de la colonne. Il est possible que tous les éléments de cette matrice ne soient pas connus (ce sont les données manquantes). On note $\mathcal{D}$, l'ensemble des couples (i, j) pour lesquels X_{ij} est connu. S'il n'y a pas de données manquantes $\mathcal{D} = [1 \dots I] \times [1 \dots J]$.

2.1.b Données Complémentaires

Il peut arriver que l'on dispose d'informations additionnelles au niveau des lignes et des colonnes de la matrice $\mathcal{X}$. Parfois, il s'agira de véritables covariables, d'autres fois cette information se réduira aux vecteurs dont toutes les composantes sont égales à l'unité (cf § III.1), qui peuvent être assimilés à des covariables indiquant que toutes les lignes [respectivement les colonnes] jouent des rôles identiques. Dans une matrice $\mathcal{X}$ de type Individus $\times$ Variables, les covariables pourraient être des caractéristiques des Individus comme l'âge, le sexe, la catégorie socio-professionnelle...

La matrice dont les lignes correspondent aux lignes de $\mathcal{X}$ et dont les colonnes correspondent aux covariables associées aux lignes de $\mathcal{X}$ sera notée $\mathcal{Y}$, de terme courant Y_{ik} avec $i \in [1 \dots I]$, $k \in [1 \dots K]$. Contrairement à $\mathcal{X}$, cette matrice est supposée connue complètement (sans données manquantes).

Symétriquement, la matrice dont les lignes correspondent aux colonnes de $\mathcal{X}$ et dont les colonnes correspondent aux covariables associées aux colonnes de $\mathcal{X}$ sera notée $\mathcal{Z}$, de terme courant Z_{jh} avec $j \in [1 \dots J]$, $h \in [1 \dots H]$. Tout comme $\mathcal{Y}$, cette matrice est supposée connue complètement (sans données manquantes).

De chacune des deux matrices $\mathcal{Y}$ et $\mathcal{Z}$, sont extraites deux sous matrices (on verra par la suite qu'elles doivent être disjointes) dont deux seront utilisées dans la

partie linéaire : $\mathcal{Y}^1$ et $\mathcal{Z}^1$, les deux autres pour la partie bilinéaire du modèle : $\mathcal{Y}^b$ et $\mathcal{Z}^b$. Ces 4 matrices sont constituées par une sélection de sous-ensembles des colonnes des matrices originales. Leurs dimensions respectives seront :

$$\mathcal{Y}^1[I, K^1]; \mathcal{Z}^1[J, H^1]; \mathcal{Y}^b[I, K^b]; \mathcal{Z}^b[J, H^b]$$

On supposera, sans perte de généralité, que ces 4 ensembles de covariables ne sont pas colinéaires, c'est à dire que les matrices sont de plein rang colonne.

2.1.c Divers

La matrice identité de dimension N sera notée $\mathbb{I}_N$; le vecteur de dimension N dont toutes les composantes sont unité sera noté $\mathbb{1}_N$; la transposée de la matrice $\mathcal{M}$ sera notée $\mathcal{M}'$. Si A est une variable aléatoire, son espérance mathématique sera notée $\mathcal{E}[A]$.

2.2 Définition Générale

Par «modèle» ou «approximation» linéaire et bilinéaire sous contrainte linéaire, on entend l'ajustement de la matrice de données $\mathcal{X}$ à la forme suivante :

$$\underset{[K^1, H^1]}{\mathcal{Y}^1 \cdot \mu \cdot \mathcal{Z}^{1'}} + \underset{[I, H^1]}{\alpha \cdot \mathcal{Z}^{1'}} + \underset{[K^1, J]}{\mathcal{Y}^1 \cdot \beta'} + \underset{[K^b, R][R, R][R, H^b]}{\mathcal{Y}^b \cdot \gamma \cdot \theta \cdot \delta' \cdot \mathcal{Z}^{b'}} \quad (\text{MG})$$

où les lettres grecques représentent des matrices de paramètres à ajuster dont les dimensions respectives sont données entre crochets. La matrice θ est une matrice diagonale. Les matrices μ et θ , parfaitement redondantes, ont été introduites pour garder la symétrie entre le rôle joué par les lignes et celui joué par les colonnes.

Les 3 premiers termes sont linéaires en paramètres, le dernier est bilinéaire en fonction des paramètres (θ est un paramètre fantôme qui peut être intégré à γ comme à δ).

3. Relations avec les modèles classiques

La forme (MG) recouvre une grande variété de formules utilisées dans le traitement statistique des données, que ce soit dans le cadre de l'analyse de variance entre deux facteurs ou dans ce celui de l'analyse des données. Dans cette section sont présentées les plus classiques; la discussion générale sur les paramètres et la justification des dimensions paramétriques sont renvoyées en §IV et §V.

3.1 Modèle Additif

C'est le modèle de base utilisé en analyse de variance. La réponse aux deux facteurs est, en espérance, la somme de deux fonctions : l'une ne dépendant que des lignes (premier facteur), l'autre ne dépendant que des colonnes (second facteur) :

$$\mathcal{E}[X_{ij}] = \mu + \alpha_i + \beta_j$$

ce qui matriciellement donne :

$$\underset{[1,1]}{\mathbb{1}_I} \cdot \underset{[1,1]}{\mu} \cdot \underset{[1,J]}{\mathbb{1}'_J} + \underset{[I,1]}{\alpha} \cdot \underset{[I,1]}{\mathbb{1}'_J} + \underset{[1,J]}{\mathbb{1}_I} \cdot \underset{[1,J]}{\beta'} \quad (\text{ADD})$$

Dans ce modèle, les covariables sont réduites aux vecteurs $\mathbb{1}_I$ et $\mathbb{1}_J$, il ne contient pas de termes bilinéaires, la dimension paramétrique est $(I + J - 1)$.

3.2 Modèle Multiplicatif

C'est le pendant du modèle additif lorsque les 2 fonctions de chacun des 2 facteurs (ligne et colonne) interviennent multiplicativement et non plus additivement. Les 2 modèles ont été confrontés par Fisher et Mackenzie (1923). L'avantage revient en général, au modèle additif. La raison probable est qu'il est invariant par translation, qualité qui dans certains types de données est essentielle. D'autre part son traitement est beaucoup plus simple, puisque dans le cas de données complètes, il ne fait intervenir que des sommes contre une diagonalisation de matrice dans le cas du modèle multiplicatif (cf §VI). On peut l'écrire :

$$\mathcal{E}[X_{ij}] = \theta \cdot \gamma_i \cdot \delta_j \quad (\text{mul})$$

ce qui matriciellement et en le généralisant à un nombre quelconque de termes multiplicatifs donne :

$$\underset{[I,R][R,R][R,J]}{\mathbb{1}_I \cdot \gamma \cdot \theta \cdot \delta' \cdot \mathbb{1}'_J} \quad (\text{MUL})$$

Il ne contient que des termes bilinéaires, les covariables considérées sont «maximales» dans le sens où le rajout de n'importe quelle covariable n'augmenterait pas le modèle. La dimension paramétrique de (mul) est $(I + J - 1)$, celle de (MUL) est $R(I + J - R)$.

3.3 Modèles de Régression Factorielle

Ces modèles ont été utilisés par plusieurs auteurs, en général dans des contextes appliqués particuliers. Citons entre autres Hardwick et Wood (1972) qui n'utilisent de termes linéaires que pour une des entrées (par exemple les lignes) de la matrice à expliquer. Gabriel (1978), du point de vue de l'approximation de matrices, a proposé ce genre de modèles en leur adjoignant une partie multiplicative. Denis (1980, 1983) en a fait une exploration systématique considérant ces modèles comme une extension directe du modèle additif (ADD) et leur a donné l'appellation de *régression factorielle*.

$$\mathcal{E}[X_{ij}] = \sum_h \sum_k Y_{ik} \cdot \mu_{kh} \cdot Z_{jh} + \sum_h \alpha_{ih} \cdot Z_{jh} + \sum_k Y_{ik} \cdot \beta_{jk}$$

ce qui matriciellement correspond à la partie linéaire de (MG) :

$$\underset{[K,H]}{\mathcal{Y}} \cdot \underset{[I,H]}{\mu} \cdot \underset{[K,J]}{\mathcal{Z}'} + \underset{[I,H]}{\alpha} \cdot \underset{[K,J]}{\mathcal{Z}'} + \underset{[K,J]}{\mathcal{Y}} \cdot \underset{[K,J]}{\beta'} \quad (\text{RF})$$

La dimension paramétrique de ce modèle est $(I \cdot H + J \cdot K - KH)$.

3.4 Analyse en Composantes Principales non Normée

Nous quittons la classe des modèles issus de l'école inférentielle et abordons les approximations utilisées en *Analyse des Données à la Française*. Il est clair que l'Analyse des Données ne se réduit pas à notre présentation et que nous ne la considérons que d'un point de vue très particulier, celui de l'approximation d'une matrice par une matrice d'un rang inférieur.

Les idées d'Analyse en Composantes Principales sont anciennes puisqu'on cite souvent Pearson (1901) qui posa le problème en terme d'ajustements géométriques. Eckart et Young (1936) traitèrent le problème dans le contexte de l'approximation de matrices. Plus récemment Cailliez et Pagès (1976) en donnèrent une vision très cohérente, regroupant dans une même classe grâce au schéma de dualité toutes une famille de méthodes. Si R facteurs sont retenus, cette approximation peut s'écrire :

$$\beta_j + \sum_{r=1}^R \theta^r \cdot \gamma_i^r \cdot \delta_j^r$$

où β_j correspond au centrage préalable des colonnes, ce qui matriciellement donne :

$$\underset{[1,J]}{\mathbb{1}_I \cdot \beta'} + \underset{[I,R]}{\mathbb{1}_I \cdot \gamma} \cdot \underset{[R,R]}{\theta} \cdot \underset{[R,J]}{\delta'} \cdot \underset{[1,J]}{\mathbb{1}'_J} \quad (\text{ACP})$$

La dimension paramétrique de cette approximation est $J + R((I - 1) + J - R)$.

3.5 Modèle à Interaction Multiplicative

Un modèle englobant à la fois le modèle additif (ADD) et le modèle multiplicatif (MUL) a été développé par plusieurs auteurs (Gollob, 1968 ; Mandel, 1971 ; Johnson et Graybill, 1972 ; Corsten et van Eijnsbergen, 1972) dans le contexte de l'étude de l'interaction entre deux facteurs :

$$\mathcal{E}[X_{ij}] = \mu + \alpha_i + \beta_j + \sum_{r=1}^R \gamma_i^r \cdot \theta^r \cdot \delta_j^r$$

ce qui matriciellement s'écrit :

$$\underset{[1,1]}{\mathbb{1}_I \cdot \mu \cdot \mathbb{1}'_J} + \underset{[I,1]}{\alpha \cdot \mathbb{1}'_J} + \underset{[1,J]}{\mathbb{1}_I \cdot \beta'} + \underset{[I,R]}{\mathbb{1}_I \cdot \gamma} \cdot \underset{[R,R]}{\theta} \cdot \underset{[R,J]}{\delta'} \cdot \underset{[1,J]}{\mathbb{1}'_J} \quad (\text{IMU})$$

La dimension paramétrique de ce modèle est $(I + J - 1) + R((I - 1) + (J - 1) - R)$.

Ce modèle est très proche du précédent (ACP) puisqu'il n'en diffère que par l'ajout du terme en α (le terme en μ est implicitement compris dans le terme en β). Contrairement à l'ACP, ces modèles d'interaction multiplicative sont utilisés avec R relativement petit (généralement 1, très rarement 3 ou plus).

3.6 Analyse en Composantes Principales sous contraintes linéaires

Il s'agit d'approximation en forme (ACP) mais en imposant une contrainte linéaire à une des deux séries des composantes principales (par exemple celles correspondant aux lignes de la matrice). Les contraintes linéaires sont introduites en imposant à la série des composantes principales d'être combinaisons linéaires de variables explicatives généralement nommées *variables instrumentales*. La référence initiale semble être celle de Rao (1964); Obadia (1978) a retrouvé une démarche analogue, reprise plus tard dans divers travaux étendant les résultats à d'autres analyses factorielles (Sabatier *et al.*, 1989; Fraile *et al.*, 1989). Dans notre présentation limitée à l'ACP, l'approximation est de la forme :

$$\beta_j + \sum_{r=1}^R \theta^r \cdot \left[\sum_{k=1}^k Y_{ik} \cdot \gamma_k^r \right] \cdot \delta_j^r$$

ce qui matriciellement donne :

$$\underset{[1,J]}{\mathbb{1}_I} \cdot \beta' + \underset{[K,R]}{\mathcal{Y}} \cdot \underset{[R,R]}{\gamma} \cdot \underset{[R,J]}{\theta} \cdot \underset{[R,J]}{\delta'} \cdot \underset{[R,J]}{\mathbb{1}_j'} \quad (\text{ACPI})$$

la dimension paramétrique de l'approximation est $J + R(K + J - R)$

3.7 Modèles d'Interaction Multiplicative sous contraintes linéaires

Citant Rao (1964) des auteurs, analysant les interactions entre deux facteurs, ont proposé également de contraindre les interactions multiplicatives par l'introduction de covariables liées à un des facteurs. C'est le cas de Wood (1976) :

$$\mathcal{E}[X_{ij}] = \mu + \alpha_i + \beta_j + \sum_{r=1}^R \left[\sum_{k=1}^K Y_{ik} \cdot \gamma_k^r \right] \cdot \theta^r \cdot \delta_j^r$$

ce qui matriciellement se traduit par :

$$\underset{[1,1]}{\mathbb{1}_I} \cdot \underset{[1,1]}{\mu} \cdot \underset{[1,1]}{\mathbb{1}_j'} + \underset{[I,1]}{\alpha} \cdot \underset{[I,1]}{\mathbb{1}_j'} + \underset{[1,J]}{\mathbb{1}_I} \cdot \underset{[1,J]}{\beta'} + \underset{[K,R]}{\mathcal{Y}} \cdot \underset{[R,R]}{\gamma} \cdot \underset{[R,J]}{\theta} \cdot \underset{[R,J]}{\delta'} \cdot \underset{[R,J]}{\mathbb{1}_j'} \quad (\text{IMU1})$$

la dimension paramétrique de ce modèle est $(I + J - 1) + R(K + (J - 1) - R)$

Une symétrisation de ce modèle sur l'autre facteur a été proposée par Denis (1983, 1988). Il s'agit d'imposer des contraintes linéaires pour l'autre facteur dans le terme multiplicatif :

$$\mathcal{E}[X_{ij}] = \mu + \alpha_i + \beta_j + \sum_{r=1}^R \left[\sum_{k=1}^K Y_{ik} \cdot \gamma_k^r \right] \cdot \theta^r \cdot \left[\sum_{h=1}^H Z_{jh} \cdot \delta_h^r \right]$$

ce qui matriciellement donne :

$$\underset{[1,1]}{\mathbb{1}_I} \cdot \underset{[I,1]}{\mu} \cdot \underset{[1,J]}{\mathbb{1}'_J} + \underset{[I,1]}{\alpha} \cdot \underset{[1,J]}{\mathbb{1}'_J} + \underset{[1,J]}{\mathbb{1}_I} \cdot \underset{[K,R]}{\beta'} + \underset{[K,R]}{\mathcal{Y}} \cdot \underset{[R,R]}{\gamma} \cdot \underset{[R,J]}{\theta} \cdot \underset{[R,J]}{\delta'} \cdot \underset{[R,J]}{\mathcal{Z}'} \quad (\text{IMU2})$$

la dimension paramétrique de ce modèle est $(I + J - 1) + R(K + H - R)$.

Le modèle général (MG), dont on propose l'étude ici, est simplement la généralisation la plus simple des deux modèles les plus complets présentés dans cette brève revue : (RF) et (IMU2). Toutes les considérations qui suivent se rapporteront donc au modèle général (MG), elles s'appliqueront comme cas particulier aux exemples énumérés ci-dessus moyennant une orthogonalisation des covariables comme il est précisé dans la section suivante.

4. Contraintes sur les paramètres et les covariables

Aucune des formes particulières présentées ci-dessus n'est définie en elle-même du fait d'une sur-paramétrisation. C'est vrai au niveau de la partie linéaire, de la partie bilinéaire et des relations entre les parties linéaire et bilinéaire. La conséquence est l'indétermination des valeurs des paramètres. L'emploi de contraintes supplémentaires permet de la lever mais aussi, si les contraintes sont bien choisies, de faciliter l'expression des estimateurs. Explicitons ces divers points dans le cas de modèles simples avant d'aborder le cas général (MG).

4.1 cas du modèle additif (ADD)

Si toutes les valeurs de $\mathcal{E}[X_{ij}]$ sont connues, on ne peut pas pour autant prétendre que les valeurs des paramètres de (ADD) sont déterminées. Effectivement si μ^0 , α_i^0 et β_j^0 sont tels que :

$$\mathcal{E}[X_{ij}] = \mu^0 + \alpha_i^0 + \beta_j^0$$

alors $\mu^1 = (\mu^0 - a - b)$, $\alpha_i^1 = (\alpha_i^0 + a)$ et $\beta_j^1 = (\beta_j^0 + b)$ vérifient aussi :

$$\mathcal{E}[X_{ij}] = \mu^1 + \alpha_i^1 + \beta_j^1$$

il y a donc une infinité de séries de paramètres redonnant les mêmes valeurs de l'espérance, et pour fixer la paramétrisation il convient d'imposer des conditions supplémentaires. Il y a une infinité de manières de les choisir; ce choix doit être guidé par l'interprétation des paramètres qui en résulte. Nous utiliserons celles qui font jouer un rôle symétrique aux deux facteurs et à leurs niveaux. Elles aboutissent au centrage des effets principaux.

$$\sum_i \alpha_i = 0 \quad \text{et} \quad \sum_j \beta_j = 0$$

$\alpha_i(\beta_j)$ est alors l'écart de la moyenne arithmétique non pondérée de la ligne i (colonne j) à la moyenne générale.

4.2 cas du modèle multiplicatif (mul)

De manière similaire le modèle multiplicatif (mul) est indéterminé :

$$\theta \cdot \gamma_i \cdot \delta_j = \frac{\theta}{a \cdot b} \cdot (a \cdot \gamma_i) \cdot (b \cdot \delta_j)$$

avec a et b supposés non nuls. Les conditions les plus usitées sont :

$$\sum_i \gamma_i^2 = 1 \quad \text{et} \quad \sum_j \delta_j^2 = 1$$

4.3 cas du modèle d'interaction multiplicative à un terme

Si maintenant on considère un modèle simple ayant une partie linéaire et une partie bilinéaire, à chacune desquelles on a imposé les contraintes précédentes, le modèle n'est pas déterminé pour autant.

$$\mathcal{E}[X_{ij}] = \mu + \alpha_i + \beta_j + \gamma_i \cdot \theta \cdot \delta_j$$

il peut en effet se récrire selon le même modèle avec des paramètres différents :

$$(\mu - c \cdot \theta \cdot d) + (\alpha_i - d \cdot \theta \cdot \gamma_i) + (\beta_j - c \cdot \theta \cdot \delta_j) + \frac{(\gamma_i + c)}{c_0} \cdot [c_0 d_0 \theta] \cdot \frac{(\delta_j + d)}{d_0}$$

avec $c_0 = [\sum (\gamma_i + c)^2]^{1/2}$ et $d_0 = [\sum (\delta_j + d)^2]^{1/2}$. Les conditions complémentaires usuelles utilisées pour déterminer complètement le modèle sont :

$$\sum_i \gamma_i = 0 \quad \text{et} \quad \sum_j \delta_j = 0$$

4.4 cas général

Une série complète de conditions pour déterminer le modèle général (MG) peut être obtenue par trois systèmes d'équations. Les deux premières concernent les paramètres :

$$\begin{aligned} \mathcal{Z}^1 \cdot \beta = 0; \mathcal{Y}^1 \cdot \alpha = 0 \quad \text{si} \quad (H^1 \neq 0) \quad \text{et} \quad (K^1 \neq 0) & \quad (\text{CLL}) \\ \gamma' \cdot \gamma = \delta' \cdot \delta = \mathbb{I}_R & \quad (\text{CBB}) \end{aligned}$$

La dernière concerne les covariables.

$$\mathcal{Z}^1 \cdot \mathcal{Z}^b = 0; \mathcal{Y}^1 \cdot \mathcal{Y}^b = 0 \quad (\text{CCV})$$

Ces conditions ne restreignent pas la généralité de (MG). (CLL) évite la redondance entre les trois termes linéaires. (CBB) impose les conditions habituelles de l'analyse

en composantes principales. (CCV) correspond au fait que si une covariable est incluse dans la partie linéaire, elle n'apporte rien de plus dans la partie bilinéaire du modèle. L'intérêt de ce système est qu'il assure, en particulier, l'orthogonalité (pour la métrique canonique $\mathbb{I}$) dans $\mathbb{R}^{IJ}$ des quatre termes de (MG).

Tels qu'ils sont écrits en § III les divers cas particuliers de (MG) ne satisfont pas à la contrainte d'orthogonalité entre covariables de la partie linéaire et covariables de la partie bilinéaire (CCV) mais il est possible de les y ramener. Par exemple dans le cas de l'analyse en composantes principales (ACP), $\mathcal{Y}^1 = \mathbb{1}$ et $\mathcal{Y}^b = \mathbb{I}_J$ et donc $\mathcal{Y}^{1'} \cdot \mathcal{Y}^b \neq 0$. Mais on peut centrer les covariables bilinéaires pour faire passer une partie du terme bilinéaire dans le terme linéaire :

$$\mathbb{1}_I \cdot [\beta + I^{-1} \mathbb{1}'_I \gamma \cdot \theta \cdot \delta'] + [\mathbb{1}_I - I^{-1} \mathbb{1}_I \mathbb{1}'_I] \cdot \gamma \cdot \theta \cdot \delta' \cdot \mathbb{1}'_J$$

Sous cette nouvelle paramétrisation la condition (CCV) est bien vérifiée mais la non colinéarité des covariables bilinéaires ne l'est plus. Il faut donc remplacer $[\mathbb{1} - I^{-1} \mathbb{1} \mathbb{1}'] \cdot \gamma$ par $\mathcal{C} \cdot \tilde{\gamma}$ où $\mathcal{C}$ est une matrice $[I, I - 1]$ dont les colonnes forment une base de $\mathbb{1}'_I$ (par exemple $[\mathbb{1}_I - I^{-1} \mathbb{1}_I \mathbb{1}'_I]$ dont on a supprimé une colonne) et $\tilde{\gamma}$ est une matrice $(I - 1) \times R$.

5. Dimensions paramétriques

Un point capital est celui de la détermination de la dimension paramétrique. Celle-ci se fait aisément en utilisant une représentation du modèle (ou de l'approximation) par une *table récapitulative* telle que l'a introduite Denis (1983, 1988). Cette représentation est basée sur la décomposition tensorielle des sous-espaces vectoriels de $\mathbb{R}^{IJ}$ par lesquels est décomposée la matrice $\mathcal{X}$. Elle correspond aux conditions sur les paramètres et les covariables qui ont été proposées en §IV.4.

$\mathbb{R}^I$ est décomposé en 3 sous-espaces vectoriels supplémentaires orthogonaux respectivement engendrés par les colonnes de $\mathcal{Y}^1$ (de dimension K^1), celles de $\mathcal{Y}^b$ (de dimension K^b) et leur supplémentaire orthogonal (de dimension $I - K^1 - K^b$). Il en est de même pour $\mathbb{R}^J$ avec les H^1 covariables $\mathcal{Z}^1$ et les H^b covariables $\mathcal{Z}^b$.

Ces deux décompositions croisées par produit tensoriel, créent 9 sous-espaces vectoriels de $\mathbb{R}^I \otimes \mathbb{R}^J$, base de la représentation de (MG), voir la figure 1.

$[\mathcal{Y}^1 \cdot \mu \cdot \mathcal{Z}^1]$ correspond à $\mathcal{A}$, sa dimension paramétrique est $K^1 \cdot H^1$.

$[\alpha \cdot \mathcal{Z}^1']$ correspond à $\mathcal{D} \cup \mathcal{G}$, sa dimension paramétrique est $(I - K^1) \cdot H^1$.

$[\mathcal{Y}^1 \cdot \beta']$ correspond à $\mathcal{B} \cup \mathcal{C}$, sa dimension paramétrique est $K^1 \cdot (J - H^1)$.

$[\mathcal{Y}^b \cdot \gamma \cdot \theta \cdot \delta' \cdot \mathcal{Z}^{b'}]$ correspond à une partie de $\mathcal{E}$, définie par l'approximation de $\mathcal{X} - [\mathcal{Y}^1 \cdot \mu \cdot \mathcal{Z}^{1'} + \alpha \cdot \mathcal{Z}^1 + \mathcal{Y}^1 \cdot \beta']$ par une matrice de rang R respectant les contraintes linéaires définies par les covariables bilinéaires. Sa dimension paramétrique s'obtient facilement en soustrayant les $R(R + 1)$ contraintes indépendantes de (CBB) au nombre de paramètres, ce qui donne $R \cdot (K^b + H^b - R)$. Cette décomposition

	$\{Z^1\}$ $[K^1]$	$\{Z^b\}$ $[H^b]$	$\{Z^1, Z^b\}^\perp$ $[J - H^1 - H^b]$
$\{Y^1\}, [K^1]$	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$
$\{Y^b\}, [K^b]$	$\mathcal{D}$	$\mathcal{E}$	$\mathcal{F}$
$\{Y^1, Y^b\}^\perp, [I - K^1 - K^b]$	$\mathcal{G}$	$\mathcal{H}$	$\mathcal{I}$

FIGURE 1

Dénomination des sous-tables de la table récapitulative du modèle (MG). En entrées de la tables (lignes et colonnes) se trouvent les décompositions de $\mathbf{R}^I$ et $\mathbf{R}^J$ obtenues à partir des ensembles de covariables. La dimension des sous-espaces vectoriels associés à chacune des 9 cases du tableau est donnée par le produit des dimensions de sa ligne et de sa colonne. Par exemple la dimension de $\mathcal{F}$ est $K^b \cdot (J - H^1 - H^b)$.

peut se poursuivre comme il est indiqué dans la figure 2 :

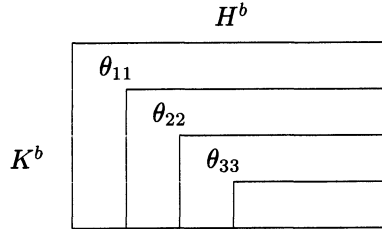


FIGURE 2

Décomposition de la case $\mathcal{E}$ de la figure 1. La dimension paramétrique du terme multiplicatif contenant θ_{11} est $K^b + H^b - 1$; celle du terme générique θ_{ii} est $K^b + H^b - (2 \cdot i - 1)$; elles sont proportionnelles aux aires des bandes ligne + colonne dans lesquelles ils sont placés.

Dans le même style de représentation géométrique, les dimensions paramétriques des termes de (MG) peuvent être récapitulées dans la figure 3 :

	$\{Z^1\}$ [H ¹]	$\{Z^b\}$ [H ^b]	$\{Z^1, Z^b\}^\perp$ [J - H ¹ - H ^b]
$\{Y^1\}$ [K ¹]	$Y^1 \cdot \mu \cdot Z^{1'}$ [H ¹ · K ¹]	$Y^1 \cdot \beta'$ [K ¹ · (J - H ¹)]	
$\{Y^b\}$ [K ^b]	$\alpha \cdot Z^{1'}$ [(I - K ¹)H ¹]	$Y^b \gamma \theta \delta' Z^{b'}$ [P]	Ecart à l'Ajustement
$\{Y^1, Y^b\}^\perp$ [I - K ¹ - K ^b]			

FIGURE 3

table récapitulative du modèle général. Dans chaque partie de la table sont rappelés les termes correspondants du modèle et entre crochets, leurs dimensions paramétriques ($P = R(K^b + H^b - R)$).

La dimension paramétrique totale de (MG) est donc :

$$I \cdot H^1 + J \cdot K^1 - H^1 \cdot K^1 + R \cdot (K^b + H^b - R) \quad (DP)$$

6. Ajustement des moindres carrés sans frontières manquantes

Si la matrice $\mathcal{X}$ que l'on veut ajuster à la forme générale (MG) est complète, c'est à dire que l'on dispose d'une observation pour chacune de cases, il est possible d'exhiber explicitement les estimations des *moindres carrés* de tous les paramètres, valeurs des paramètres qui minimisent la quantité :

$$\sum_{i=1}^I \sum_{j=1}^J [X_{ij} - \mathcal{E}[X_{ij}]]^2 = \|\mathcal{X} - [Y^1 \mu Z^{1'} + \alpha Z^{1'} + Y^1 \beta' + Y^b \gamma \theta \delta' \cdot Z^{b'}]\|^2$$

Avec les contraintes supplémentaires (CLL), (CBB) et (CCV), l'annulation des dérivées premières conduit rapidement aux équations suivantes (le $\hat{\cdot}$ indique qu'il s'agit de l'estimateur) :

$$\hat{\mu} = [Y^{1'} \cdot Y^1]^{-1} Y^{1'} \cdot \mathcal{X} \cdot Z^1 [Z^{1'} \cdot Z^1]^{-1} \quad (EL1)$$

$$\hat{\alpha} = [I_I - Y^1 [Y^{1'} \cdot Y^1]^{-1} Y^{1'}] \cdot \mathcal{X} \cdot Z^1 [Z^{1'} \cdot Z^1]^{-1} \quad (EL2)$$

$$\hat{B}' = [Y^{1'} \cdot Y^1]^{-1} Y^{1'} \cdot \mathcal{X} \cdot [I_J - Z^1 [Z^{1'} \cdot Z^1]^{-1} Z^{1'}] \quad (EL3)$$

$$\hat{\theta}_{rr} = r\text{-ième valeur propre de } S(1) \text{ ou de } S(c) \quad (EB1)$$

$\hat{\gamma}$ matrice dont la r -ième colonne est
le r -ième vecteur propre de $S(l)$ (EB2)

$\hat{\delta}$ matrice dont la r -ième colonne est
le r -ième vecteur propre de $S(c)$ (EB3)

où $S(l)$ et $S(c)$ sont des matrices dont on peut vérifier qu'elles ont mêmes valeurs propres non-nulles :

$$S(l) = (\mathcal{Y}^b{}' \mathcal{Y}^b)^{-1} \cdot \mathcal{Y}^b{}' \cdot \mathcal{X} \cdot \mathcal{Z}^b \cdot (\mathcal{Z}^b{}' \mathcal{Z}^b)^{-1} \cdot \mathcal{Z}^b{}' \cdot \mathcal{X}' \cdot \mathcal{Y}^b$$

$$S(c) = (\mathcal{Z}^b{}' \mathcal{Z}^b)^{-1} \cdot \mathcal{Z}^b{}' \cdot \mathcal{X}' \cdot \mathcal{Y}^b \cdot (\mathcal{Y}^b{}' \mathcal{Y}^b)^{-1} \cdot \mathcal{Y}^b{}' \cdot \mathcal{X} \cdot \mathcal{Z}^b$$

remarques :

i) : La supposition de non-colinéarité des covariables permet d'affirmer que les inverses indiquées existent bien.

ii) : Dans les formules (EL1), (EL2) et (EL3) qui donnent les estimateurs des paramètres de la partie linéaire, les covariables bilinéaires n'interviennent pas. L'estimation de la partie linéaire est donc indépendante de la partie bilinéaire. Ceci conduit à décomposer le processus d'estimation en 2 étapes : l'estimation de la partie linéaire puis sur les résidus obtenus l'estimation de la partie bilinéaire (cf §VII.4).

iii) : Les formules ci-dessus sont bien celles de Gabriel (1978) lorsqu'il n'y a pas de contraintes linéaires sur la partie bilinéaire.

iv) : On peut vérifier que pour les cas particuliers de (MG) présentés en §III, l'application de ces équations redonne bien les formules classiques si on prend la précaution de satisfaire aux conditions supplémentaires introduites en §IV.

7. Ajustement des moindres carrés avec données manquantes

La question abordée maintenant est plus difficile, car on ne suppose plus la matrice $\mathcal{X}$ complètement connue. Dans le cadre de l'analyse de variance, on dit être dans un cas «non-orthogonal», expression qui traduit, entre autres, le fait que toutes les formules précédentes ne s'écrivent plus commodément par des produits matriciels. Si on reprend l'exemple du modèle additif (ADD), les estimateurs ne peuvent plus être donnés par une formule explicite contrairement au cas orthogonal.

7.1 Objectif

Le problème est de chercher les estimations des moindres carrés, valeurs des paramètres $(\hat{\mu}, \hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\theta} \text{ et } \hat{\delta})$ qui minimisent la somme des carrés résiduelle :

$$\sum_{(i,j) \in \mathcal{D}} [X_{ij} - \mathcal{E}[X_{ij}]]^2 \quad (\text{CRI})$$

où $\mathcal{D}$ est l'ensemble des couples (i, j) pour lesquels la valeur X_{ij} est connue.

7.2 L'Idée Générale

La proposition d'estimation est en fait une heuristique algorithmique employée avec succès sur des lots de données divers depuis une année. Il s'agit de l'application de la technique des *Moindres Carrés Alternés* (Alternating Least Squares) dont une description et une utilisation dans un autre contexte sont présentées dans Kroonenberg (1983).

L'algorithme est basé sur la remarque élémentaire que conditionnellement aux paramètres dépendants du facteur ligne (i), le modèle est linéaire en paramètres et réciproquement, conditionnellement aux paramètres dépendants du facteur colonne (j), le modèle est linéaire en paramètres. L'idée est donc de choisir des valeurs initiales pour les paramètres α et γ et, en les supposant exacts, d'estimer les paramètres μ , β , θ et δ ; puis de recommencer dans l'autre sens en supposant cette fois-ci les valeurs de β et δ connues.

Il convient de faire remarquer que si la conception de l'algorithme est relativement simple, sa mise en oeuvre pose un certain nombre de difficultés informatiques dues en particulier à l'obligation de manier des matrices comprenant des données manquantes. Une réalisation en TurboPascal a été faite dans le logiciel INTERA (Decoux et Denis, 1990) dont l'expérimentation est la base des idées présentées ici.

7.3 Convergence de l'Algorithme

A chaque itération la valeur du critère (CRI) ne peut que diminuer; puisqu'elle est bornée par 0 la convergence est assurée. Cependant, il est clair que la convergence pourrait se faire vers un minimum local. Dans le cas du modèle multiplicatif (mul) et de données complètes, si on prend comme valeur de départ de γ un vecteur propre de $X \cdot X'$ ne correspondant pas à la plus grande valeur propre, la convergence sera immédiate mais le maximum global ne sera pas atteint. Cependant la probabilité d'initialiser avec un tel vecteur est presque nulle et ce minimum local n'est pas stable numériquement. Il faut donc choisir soigneusement les valeurs initiales.

7.4 Initialisation

Comme il vient d'être mentionné, ce peut être un point capital et de fait, l'expérience montre que suivant le point de départ de l'algorithme la convergence est lente ou rapide. La stratégie suivante est utilisée dans INTERA.

Tout d'abord, l'ajustement de la seule partie linéaire est réalisée. On peut par exemple utiliser les résultats du modèle linéaire pour l'obtenir. A partir de cet ajustement, les résidus de la partie linéaire sont calculés et utilisés pour ajuster la partie bilinéaire du modèle (grâce à l'algorithme présenté). Dans le cas orthogonal (sans données manquantes), Kruskal (1977) puis Gabriel (1978) ont démontré que cette procédure en deux étapes fournissait l'estimation globale. On vérifie numériquement que c'est d'autant moins vrai que les données manquantes sont nombreuses. Néanmoins ces deux étapes préliminaires aux dimensions paramétriques allégées

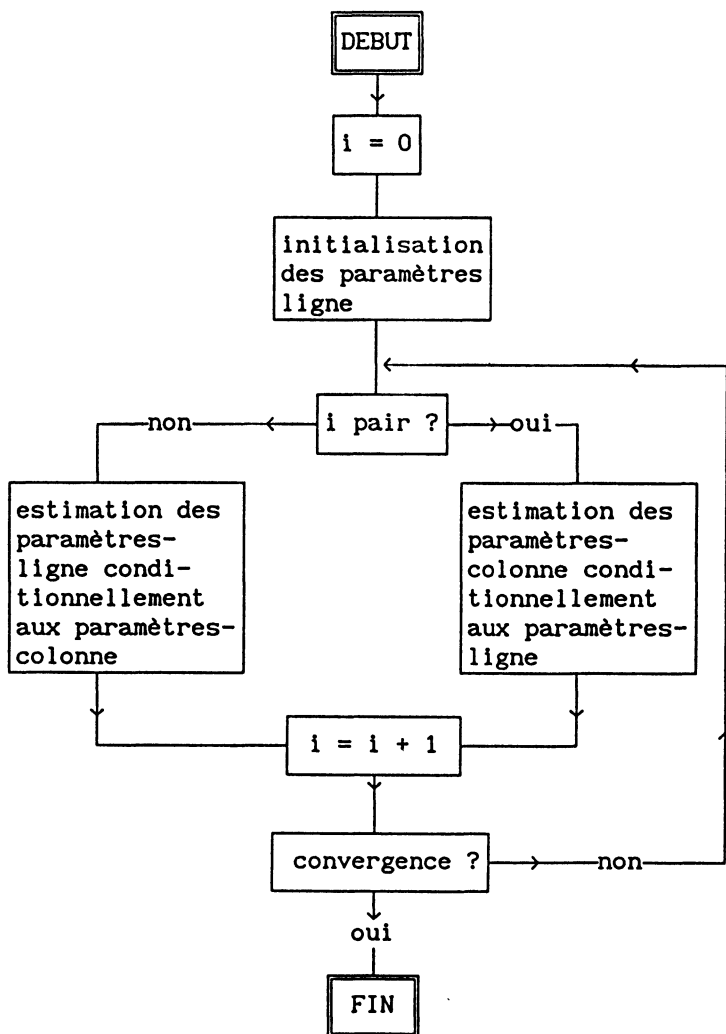


FIGURE 4

*diagramme général de l'estimation du modèle (MG)
dans le cas de données manquantes par l'algorithme MCA.*

donnent de bonnes valeurs de départ pour l'application de l'algorithme sur l'ajustement global tel qu'il a été présenté en §VII.2.

7.5 Test d'Arrêt

Différents tests d'arrêt ont été essayés, en particulier basés sur la stabilité des valeurs des paramètres estimés à chaque étape. Finalement, c'est le critère le plus simple qui a été retenu. Il consiste à observer la diminution du critère (CRI) et à stopper le processus lorsque le gain est inférieur à une proportion donnée du critère. Si on note $CRI[i]$ la valeur du critère à l'étape i et ε une valeur petite (par exemple 10^{-5}) déterminée à l'avance, l'arrêt est conditionné par la réalisation de l'inégalité :

$$CRI[i - 1] - CRI[i] < \varepsilon \cdot CRI[i - 1] \quad (\text{ARE})$$

Il peut arriver, contrairement à la théorie, que $CRI[i]$ soit supérieur à $CRI[i - 1]$. C'est l'indication que ε est trop faible par rapport à la précision numérique du processus. Ceci arrive d'autant plus fréquemment que la taille de la matrice est importante.

7.6 Comparaison avec l'algorithme E.M.

Un algorithme très général d'estimation en cas de données manquantes est celui dit «EM» pour *Espérance-Maximisation* (Expectation-Maximization) dont une excellente présentation est faite par Dempster, Laird et Rubin (1977). L'idée de base est d'estimer progressivement les données manquantes pour pouvoir appliquer une procédure d'estimation habituelle. Il s'agit d'un algorithme itératif dont chaque itération est composée de 2 étapes : *Espérance* et *Maximisation* de la Vraisemblance. Appliqué au cas du modèle (MG) il se réduit à une suite de calculs très faciles à mettre en œuvre. L'initialisation pourrait consister à remplacer les cases de $\mathcal{X}$ inconnues par la moyenne générale et cette matrice complétée pourrait fournir par l'application des formules données en §VI des valeurs initiales des paramètres du modèle. Ensuite les étapes d'*Espérance* (remplacement des cases inconnues par l'espérance donnée par les paramètres de l'étape précédente) et de *Maximisation* (estimation des paramètres par la méthode du Maximum de Vraisemblance, §VI) peuvent se succéder jusqu'à convergence d'un critère semblable à (ARE).

Malgré la réputation et la simplicité de cet algorithme, trois raisons militent pour que l'algorithme MCA lui soit préféré :

i) : Quelques comparaisons ont montré qu'une itération de MCA prenait beaucoup moins de temps qu'une itération de EM alors qu'à des points de départ équivalents, elle était toujours légèrement plus efficace. La rapidité se comprend par le fait que MCA ne pratique que l'inversion d'une matrice alors qu'EM demande la diagonalisation d'une matrice carrée symétrique.

ii) : S'il n'y a qu'une ou zéro observation par case de $\mathcal{X}$, la notion de plan complet est évident : c'est une observation par case, mais en cas de répétitions (comme le permet INTERA), la définition de plan orthogonal associé est plus problématique.

iii) : Enfin en terme d'occupation mémoire, EM est beaucoup plus gourmand que MCA.

8. Problème de l'estimabilité

La possibilité de pratiquer des ajustements sur des données comprenant des données manquantes ouvre la voie à de nombreux problèmes intéressants. Un de ceux-ci est la question de l'estimabilité ou identifiabilité du modèle. Si il y a trop de données manquantes, le modèle peut être trop riche pour pouvoir être approché par les données. Le cas trivial est celui où le nombre de données disponibles est inférieur à la dimension paramétrique ! Mais la plupart du temps la détection n'est pas si évidente.

L'algorithme proposé en §VII est basé sur l'application répétée de l'ajustement de modèles linéaires donc d'inversions de matrices. Si ces matrices ne sont pas inversibles (mais on ne dispose que d'une détection numérique), c'est souvent l'indication de non-estimabilité. Mais si cette condition est nécessaire, elle n'est pas suffisante comme on peut le vérifier avec le modèle additif (ADD) appliqué à une matrice (3,3) dont seulement 4 valeurs (alors que la dimension paramétrique du modèle est de 5) sont connues comme dans le cas de la figure 5.

	3a	3b
3c		
3d		

FIGURE 5

matrice X ne comprenant que 4 valeurs.

L'application de l'algorithme MCA sur une telle table fournit les estimations suivantes pour le modèle additif (ADD) :

$$\begin{aligned}
 \mu &= a + b + c + d - x \\
 \alpha_1 &= 2x - c - d & \beta_1 &= 2x - a - b \\
 \alpha_2 &= 2c - x - d & \beta_2 &= 2a - x - b \\
 \alpha_3 &= 2d - x - c & \beta_3 &= 2b - x - a
 \end{aligned}$$

qui si elles ajustent exactement les valeurs de $\mathcal{X}$ et vérifient les contraintes supplémentaires (CLL) n'en sont pas moins indéterminées puisqu'elles dépendent de x .

Dans le cas général des formules de type (MG), beaucoup de travail reste à faire pour préciser les conditions de l'estimabilité.

9. Exemple

Les données utilisées sont celles des poissons d'Amiard telles que les ont retranscrites Cailliez et Pagès (1976) page 278 de leur ouvrage. Pour illustrer la procédure d'estimation avec données manquantes, 29 valeurs ont été ôtées sur les

368 que comprend la matrice $\mathcal{X}$ initiale. Les données utilisées sont reproduites dans le tableau 1.

	C.1	C.2	C.3	C.4	C.5	C.6	C.7	C.8	C.9	C.10	C.11	C.12	C.13	C.14	C.15	C.16
I.1	10	65	65	107	7	76	16	142	1	132		197	54	47	18	11
I.2	9	33	39	67	29	113	10	99	2	122	220	198	49	44	16	10
I.3	6	47	71	95	11	192	9	121	2	129	220	198	49	15		11
I.4	7	70	40	66	8	310	10	90	2	133	225	199	52	48	15	11
I.5	8	59	67	100	14		4	244	1	57	168	149	37	37	9	9
I.6	8	46	55	112	17	115	8	153	1	59	178	160	38	35	11	9
I.7	7	47	36	87	16		4	162	1	59	176	156		36	11	9
I.8	11	79	46	95	20		10	141	4	47	176	165	39	31	10	8
I.9	13	80	64	155	42		9	169	3	72	182		40	39	12	
I.10	21	150	115	146	49	229		233	5	79	200	179	45	38	12	9
I.11	12	91	84	138	22	590		220	2	80	185	163	43	41	12	11
I.12	14	120	76	125	21	309		617	5	72	175	158	40	39	13	10
I.13	14	142	86	135	34	523		211	11	75	189	169	42	39		10
I.14	23	92	80	132	49	459		197	2	52	164	147	36	35	12	9
I.15	13	85	64	124	20	318	9		4	86	195	175	41	39	16	10
I.16	14	106		110	31	115	9		6	87	210	170	46		17	10
I.18	32	224	260	314	36	107	13		3	72	181	164	41	36	13	9
I.19	22	162	218	318	25	884	5		2	63	175	160	38	35	12	9
I.20	31	195	208		73	109	5		11	49	170	154	39	33	12	8
I.21	15	127	119	197	23	99	7	157		107	204	185	47	45	15	11
I.22	22	160	256	282	12	102	11	690		83	190	176	42	44	14	9
I.23	24	162	231	308	51		17	558		82	194	168	42	39	14	
I.24	19	64	163	229	16	109	8	345		91	190	172	44	42	13	11

TABLEAU 1

données de l'exemple, les données manquantes sont figurées par une case vide ; par exemple pour («I19», «C.8»). Ce tableau ne comprend que 23 lignes car l'individu «I17» n'existe pas.

Le modèle utilisé pour ajuster ces données est le modèle (IMU) : ajustement additif et un terme multiplicatif car aucune covariable associées aux facteurs n'est disponible.

Les valeurs du critère (CRI) sont résumées dans le tableau 2 pour différentes approximations emboîtées. Les valeurs des estimations des paramètres sont reproduites dans le tableau 3.

Formules d'approximation	(CRI)	Dimension paramétrique
$X_{ij} \simeq 0$	6 908 445	0
$X_{ij} \simeq \mu$	4 093 536	1
$X_{ij} \simeq \mu + \alpha_i + \beta_j$	1 547 047	38
$X_{ij} \simeq \mu + \alpha_i + \beta_j + \gamma_i \cdot \theta \cdot \delta_j$	592 370	74

TABLEAU 2

Réduction du critère pour une série d'approximations

$$\hat{\mu} = 93.3150 \quad \text{et} \quad \hat{\theta} = 1204.5177$$

	$\hat{\alpha}_i$	$\hat{\gamma}_i$		$\hat{\beta}_j$	$\hat{\delta}_j$
I.1	-23.8205	0.0471	C.1	-77.8803	-0.0071
I.2	-27.0650	-0.0044	C.2	11.2937	0.0375
I.3	-20.2072	-0.0364	C.3	18.4709	0.0983
I.4	-12.9400	-0.1303	C.4	64.3257	0.0553
I.5	-18.0273	0.0008	C.5	-66.0977	-0.0053
I.6	-30.5025	0.0288	C.6	167.0382	-0.7469
I.7	-20.6013	-0.1041	C.7	-80.7760	-0.0169
I.8	-15.4236	-0.1327	C.8	157.8845	0.6531
I.9	-5.8504	-0.1031	C.9	-85.8353	0.0123
I10	1.9525	0.0104	C10	-11.2281	-0.0063
I11	14.5388	-0.2251	C11	94.9855	-0.0100
I12	20.6768	0.1643	C12	77.4859	-0.0118
I13	15.7359	-0.1871	C13	-51.2705	-0.0126
I14	0.7787	-0.1563	C14	-55.4689	-0.0134
I15	-7.0367	-0.0797	C15	-80.1213	-0.0169
I16	-1.8986	0.1536	C16	-82.8065	-0.0093
I18	29.8264	0.2342			
I19	26.7047	-0.6173			
I20	12.9458	0.1974			
I21	-8.4533	0.0515			
I22	40.8468	0.3557			
I23	25.6419	0.3817			
I24	2.1781	0.1509			

TABLEAU 3

Estimations des paramètres du modèle (IMU) sur les données du tableau 1, soumises aux conditions supplémentaires (CLL), (CBB) et (CLB).

10. Conclusion

Les deux propositions émises dans cet article du point de vue de l'application des méthodes statistiques concernent l'emploi de covariables et le traitement des données manquantes. Ces deux points sont cruciaux pour les biologistes utilisateurs de la statistique. Ils correspondent pour eux à la prise en compte des informations additionnelles et à l'utilisation maximale des données dont ils disposent. Néanmoins, il faut être conscient que la possibilité algorithmique de traiter des données incomplètes n'autorise pas n'importe quel cas de figure. Pour mettre en évidence la limite de ces outils, la question posée au statisticien est le calcul des variances (ou de régions de confiance) à associer à celui des estimations. Quelques premiers résultats apparaissent dans le cas de modèles simples et de données complètes (Chadœuf et Denis, 1988; Goodman and Haberman, 1990). Sans doute les capacités prodigieuses des calculateurs actuels jointes aux techniques de réchantillonnage permettront de répondre prochainement de manière opérationnelle dans le cas de données manquantes ?

Remerciements

La version définitive de cet article doit beaucoup au travail des relecteurs qui ont fait des propositions intéressantes, mis en évidence plusieurs points ambigus et amélioré la forme. Qu'ils en soient sincèrement remerciés.

Références Bibliographiques

- [1] CAILLIEZ F et PAGES J.P. (1976) Introduction à l'Analyse des Données. S.M.A.S.H., Paris, 616pp.
- [2] CHADOEUF J. et DENIS J.B. (1988) Variances Asymptotiques des Estimateurs du Terme Multiplicatif de l'Interaction et leur Utilisation. Document interne du laboratoire de Biométrie, INRA, F78026 Versailles, 47p.
- [3] CORSTEN L.C.A and VAN EIJSBERGEN A.C. (1972) Multiplicative effects in two-way analysis of variance. *Statistica Neerlandica* 26, 61-67.
- [4] DECOUX G. et DENIS J.B. (1990) INTERA, manuel d'utilisation, laboratoire de Biométrie, INRA, F78026 Versailles, 120p.
- [5] DEMPSTER A.P., LAIRD N.M. and RUBIN D.B. (1977) Maximum Likelihood from Incomplete Data via the *EM* Algorithm. *J.R.Statist.Soc.B*, 39(1),1-38.
- [6] DENIS J.B. (1980) Analyse de Régression Factorielle. *Biométrie- Praximétrie*, 20, 1-34.
- [7] DENIS J.B. (1983) Interaction entre deux facteurs. Thèse de l'Institut National Agronomique Paris-Grignon, 223p.
- [8] DENIS J.B. Denis (1988) Two-Way Analysis using Covariates. *Statistics*, 19, 123-132.
- [9] ECKART C and YOUNG G (1936) The Approximation of one Matrix by another of lower Rank. *Psychometrika*, 1(3), 211-218
- [10] FRAILE L., ESCOFFIER Y. et RAIBAUT A. (1989) Application de l'analyse des correspondances de données planifiées à l'étude de la Chemotaxie d'une larve infestante d'un copépode parasite de poisson. Soumis à Publication dans *Biometrics*.
- [11] FISHER R.A. and MACKENZIE W.A. (1923) Studies in Crop Variation II. The manurial response of different potato varieties. *Journal of Agricultural Science*, XIII(III), 311-320.
- [12] GABRIEL K.R. (1978) Least Squares Approximation of Matrices by Additive and Multiplicative Models. *J.R.Statist.Soc.B*, 40(2), 186-196.
- [13] GOLLOB H.F. (1968) A statistical model which combines features of factor analytic and analysis of variance techniques. *Psychometrika* 33, 73-115.
- [14] GOODMAN L.A. and HABERMAN S.J. (1990) The Analysis of Nonadditivity in Two-Way Analysis of Variance. *J.A.S.A.*, 85, 139-145.

- [15] GOWER J.C. and DIGBY P.G.N. (1981) Expressing Complex Relationships in Two Dimensions in Interpreting Multivariate Data, chapter 6 edited by Vic Barnett, John Wiley, 374pp.
- [16] JOHNSON D.E. and GRAYBILL F.A. (1972) An analysis of a two-way model with interaction and no replication. *Journal of the American Statistical Association* 67, 862-868.
- [17] HARDWICK R.C. and WOOD J.T. (1972) Regression Methods for studying Genotype-environment Interactions. *Heredity*, 28, 209-222.
- [18] KROONENBERG P.M. (1983) Three-Mode Principal Component Analysis. DSWO Press, Leiden, 398pp.
- [19] KRUSKAL J.B. (1977) Some Least-Squares Theorems for Matrices and N-way Arrays. Communication aux 1ères Journées Internationales d'Analyse des Données et d'Informatique organisées par l'IRIA à Versailles.
- [20] MANDEL J. (1971) A new analysis variance model for non-additive data. *Technometrics* 13, 1-18.
- [21] OBADIA J. (1978) L'Analyse en Composantes Explicatives. *Revue de Statistique Appliquée*, XXVI(4), 5-28
- [22] PEARSON K. (1901) On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, series 6(2), 559-572
- [23] SABATIER R., LEBRETON J.D., CHESSEL D. (1989) Principal Component Analysis with instrumental variables as a tool for modelling composition data. Dans *MULTIWAY DATA ANALYSIS* édité par Coppi et Bolasco, Elsevier Science Publishers B.V. (North-Holland) pp341-352.
- [24] RAO C.R. (1964), The use and interpretation of principal component analysis in applied research. *Sankhya Serie A*, A26, 329-358.
- [25] WOOD J.T. (1976) The use of environmental variables in the interpretation of genotype-environment interaction. *Heredity*, 37(1), 1-7.