REVUE DE STATISTIQUE APPLIQUÉE

P. VUAGNAT

Contribution à l'étude de l'analyse de variance triple avec effectifs inégaux

Revue de statistique appliquée, tome 21, n° 4 (1973), p. 59-67

<http://www.numdam.org/item?id=RSA_1973__21_4_59_0>

CONTRIBUTION A L'ETUDE DE L'ANALYSE DE VARIANCE TRIPLE AVEC EFFECTIFS INÉGAUX (1)

P. VUAGNAT

Institut de Statistique Mathématique. Genève

INTRODUCTION

L'analyse de variance avec un nombre quelconque de facteurs est simple lorsque le nombre d'observations par cellule, c'est-à-dire pour les différentes combinaisons des niveaux des facteurs envisagés est constant ; par contre lorsque ce nombre varie d'une cellule à l'autre l'analyse est beaucoup plus complexe. Stevens [7] a étudié le problème dans le cas de trois facteurs et a donné une méthode par approximations successives ; il fournit la suite des calculs pour un exemple dans lequel les nombres de niveaux pour les trois facteurs valent 2, 2 et 4. Comme on peut le constater et comme Stevens le reconnaît lui-même, la méthode est longue et fastidieuse.

Actuellement l'existence des ordinateurs facilite grandement la tâche et permet d'envisager certains calculs qu'on ne songeait pas à effectuer auparavant. C'est la raison pour laquelle nous avons considéré le problème de l'analyse de variance à trois facteurs dans le cas non-orthogonal d'une façon beaucoup plus générale et avons écrit un programme qui permet d'effectuer une telle analyse pour des nombres de niveaux quelconques. Le seul inconvénient de la méthode générale est, comme on le verra, que la capacité de l'ordinateur peut être facilement dépassée lorsque les nombres des niveaux sont trop élevés. Pour remédier à cet inconvénient, nous avons programmé une méthode utilisant les matrices généralisées qui permet, dans une certaine mesure, d'effectuer une analyse lorsque les nombres des niveaux sont plus élevés.

1. – MODELE LINEAIRE ET ANALYSE DE VARIANCE TRIPLE

L'analyse de variance triple peut être considérée comme un cas particulier d'un modèle linéaire qui peut s'écrire de la façon suivante en utilisant la notation vectorielle :

$$\underline{y} = X\underline{\beta} + \underline{e} \qquad (1)$$

---

(1) Article remis le 11/9/72, révisé le 23/3/73.

où $\underline{y}$ est le vecteur des observations de dimensions (nx1), si n est le nombre total des observations, X une matrice (nxp) constituée par les valeurs prises par p variables non-aléatoires, $\underline{\beta}$ le vecteur (px1) des p paramètres inconnus et $\underline{e}$ le vecteur des valeurs prises par une variable de valeur moyenne nulle.

Dans le cas qui nous intéresse d'une analyse de variance triple avec interactions doubles et un nombre de répétitions par cellule non nécessairement constant, le modèle peut s'écrire d'une façon plus explicite :

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + \delta_{ij} + \epsilon_{ik} + \eta_{jk} + e_{ijkl} \qquad (2)$$

où :

$$\begin{aligned} i &= 1, \ldots . r \\ j &= 1, \ldots . s \\ k &= 1, \ldots . t \\ l &= 1, \ldots . n_{ijk} \end{aligned}$$

n_{ijk} étant le nombre d'observations dans la $ijk^{ème}$ cellule.

Dans ce modèle le nombre de paramètres inconnus est égal à :

$$p = r + s + t + rs + rt + st = (1 + r)(1 + s)(1 + t) - rst \qquad (3)$$

Mais tous ces paramètres ne sont pas indépendants ; on a en effet les restrictions suivantes :

$$\sum_{i=1}^{r} \alpha_i = \sum_{j=1}^{s} \beta_j = \sum_{k=1}^{t} \gamma_k = 0 \qquad (4)$$

Si l'on emploie la notation matricielle, ces restrictions peuvent être exprimées de la façon suivante :

$$L'\underline{\beta} = 0 \qquad (5)$$

où L' est une matrice (mxp), m étant le nombre de restrictions.

Il est clair que dans le cas de l'analyse de variance les matrices L' et X ont des formes particulières ; leurs éléments notamment valent 0 ou 1.

Pour calculer l'estimation b de β par la méthode des moindres carrés, il faut minimiser la fonction :

$$S(b) = (\underline{y} - X\underline{b})' (\underline{y} - X\underline{b}) \qquad (6)$$

avec la restriction :

$$L'\underline{b} = 0 \qquad (7)$$

En utilisant la méthode des multiplicateurs de Lagrange ; on obtient le système d'équations suivant :

$$X' X\underline{b} + L\underline{\lambda} = X'\underline{y} \qquad (8)$$

où $\underline{\lambda}$ est un vecteur dont les composantes sont les m multiplicateurs correspondant aux m restrictions définies par le système (7). Les deux relations (7) et (8) constituent un système de (m + p) équations à (m + p) inconnues.

$$\begin{pmatrix} X'X & L \\ L' & 0 \end{pmatrix} \begin{pmatrix} \underline{b} \\ \underline{\lambda} \end{pmatrix} = \begin{pmatrix} X'\underline{y} \\ 0 \end{pmatrix} \tag{9}$$

Alors que la matrice X'X est singulière, la matrice :

$$\begin{pmatrix} X'X & L \\ L' & 0 \end{pmatrix} = G \tag{10}$$

est régulière et théoriquement la solution du système (9) peut être facilement déterminée ; en effet on peut écrire :

$$\begin{pmatrix} \underline{b} \\ \underline{\lambda} \end{pmatrix} = \begin{pmatrix} X'X & L \\ L' & 0 \end{pmatrix}^{-1} \begin{pmatrix} X'\underline{y} \\ 0 \end{pmatrix} \tag{11}$$

Pratiquement le problème est beaucoup plus compliqué du fait des dimensions de la matrice G ; en effet les dimensions de G croissent très rapidement lorsque le nombre des niveaux de l'un ou l'autre des trois facteurs augmente. Par exemple dans le problème traité par Stevens les nombres des niveaux sont :

$$r = 2 \quad s = 4 \quad t = 4$$

Ainsi les dimensions de la matrice X'X sont données par la formule (3) :

$$p = 75 - 32 = 43$$

et la dimension m de la matrice L peut être calculée à l'aide de la relation :

$$m = 2(r + s + t).$$

Dans notre cas on obtient :

$$m = 20$$

La matrice G sera donc une matrice (63 x 63).

Comme on le voit, bien que les niveaux des trois facteurs ne soient pas très élevés, les dimensions de G sont déjà repectables ; c'est la raison pour laquelle il nous a paru intéressant de voir s'il n'était pas possible de calculer l'inverse de G à partir des inverses de ses éléments sous sa forme partitionnée donnée par la relation (10), c'est-à-dire à partir des inverses des matrices X'X, L' et L. Malheureusement la matrice X'X n'est pas une matrice régulière ; toutefois, comme nous allons le voir au paragraphe suivant, on peut tourner la difficulté en introduisant la notion d'inverse généralisée et il est alors possible d'exprimer l'inverse de G à l'aide des inverses généralisées de certaines matrices singulières dont les dimensions sont inférieures aux dimensions de G.

2. – INVERSES GENERALISEES ET SOLUTION DES EQUATIONS NORMALES

Tous les auteurs n'employant pas la même terminologie, nous allons tout d'abord rappeler quelques définitions.

Inverse généralisée

Soit A une matrice (mxn) ; une matrice A^- est dite inverse généralisée de A si et seulement si elle satisfait à la relation suivante :

$$AA^- A = A$$

Remarque : Certains auteurs appellent une telle matrice une matrice conditionnelle (conditional matrix) ou c-inverse.

Inverse de Moore-Penrose

Soit A une matrice (mxn). La matrice A^+ est dite inverse de Moore-Penrose de A si elle satisfait aux conditions suivantes :

1/ AA^+ est symétrique

2/ A^+A est symétrique

3/ $AA^+A = A$

4/ $A^+AA^+ = A^+$

Remarque : Certains auteurs appellent une telle matrice une matrice généralisée ou g-inverse.

Il est évident que toute inverse de Moore-Penrose est une inverse généralisée, mais une inverse généralisée peut ne pas être une inverse de Moore-Penrose.

On peut montrer que l'inverse de Moore-Penrose est unique alors qu'il peut y avoir plusieurs inverses généralisées. D'autre part lorsque A est régulière l'inverse de Moore-Penrose est identique à l'inverse au sens classique.

Dans nos calculs nous utiliserons toujours l'inverse de Moore-Penrose.

Chakravarti [1] donne les résultats suivants que nous emploierons par la suite :

Soit A une matrice régulière ayant la forme suivante :

$$A = \begin{pmatrix} X'X & L \\ L' & 0 \end{pmatrix}$$

où les vecteurs colonnes de L n'appartiennent pas à l'espace vectoriel engendré par les vecteurs colonnes de X'X, alors l'inverse généralisée A^- de A est donnée par l'expression :

$$A^- = \begin{pmatrix} C^- - C^- LQ^- L'C^- & C^- L - C^- LQ^- Q + C^- LQ \\ Q^- L'C^- & Q^- Q - Q^- \end{pmatrix}$$

où :

$$Q = L'C^- L$$

$$C = X'X + LL'$$

Pour notre problème nous n'avons besoin que du premier terme de cette matrice partitionnée, c'est-à-dire de :

$$A_{11} = C^- - C^- LQ^- L'C^-$$

En effet si nous appliquons ce résultat à la matrice G, nous voyons que le vecteur solution $\underline{b}$ donné par la relation (11) peut être calculé de la façon suivante :

$$\underline{b} = A_{11} X'\underline{y}$$

Le calcul de A_{11} s'effectue à partir de matrices dont les dimensions sont (pxp), (mxp) et (mxm) ce qui permet une sensible économie de place lorsque on utilise un ordinateur.

Pour calculer les matrices C^- et Q^- nous avons utilisé une méthode donnée par Greville [4] qui permet de calculer l'inverse de Moore-Penrose ; avec les notations données plus haut, nous avons en fait calculé la matrice :

$$A_{11} = C^+ - C^+ LQ^+ L'C^+$$

3. – ANALYSE DE VARIANCE ET TESTS DES INTERACTIONS ET DES EFFETS PRINCIPAUX

Le modèle que nous avons choisi est celui donné par l'expression (2) qui fait intervenir les effets principaux et les interactions doubles.

Considérons tout d'abord le test suivant :

Test de l'interaction triple

Bien que l'interaction triple n'ait pas été incluse dans le modèle (2), il est possible de tester l'absence de cette interaction car il n'est pas nécessaire de connaître explicitement les estimations des paramètres qui la représentent. En effet si nous utilisons la notation vectorielle, la somme des carrés due à l'ajustement de $\underline{\beta}$ est égale à :

$$\underline{b}'X'\underline{y}$$

Comme la somme des carrés totale vaut :

$$\underline{y}'\underline{y}$$

La somme des carrés résiduelle est égale à :

$$\underline{y}'\underline{y} - \underline{b}'X'\underline{y}$$

D'autre part, comme nous avons supposé qu'il y a au moins une cellule avec plus d'une observation, nous pouvons calculer une somme de carrés à l'intérieur des cellules qui vaut :

$$\sum_{i,j,k,l} (y_{ijkl} - \bar{y}_{ijk.})^2 = \Sigma$$

où :

$$\bar{y}_{ijk.} = \frac{1}{n_{ijk}} \sum_{l=1}^{\cap ijk} y_{ijkl} \qquad \text{pour } n_{ijk} \neq 0$$

Par conséquent, nous avons l'analyse de variance suivante qui permet de tester l'interaction triple :

Source de variabilité	SC	DL	CM	F
$\underline{b}$	$\underline{b}'X'\underline{y}$	$p - m$		
Interaction	$\underline{y}'\underline{y} - \underline{b}'X'\underline{y} - \Sigma$	$n - p + m - u$	I_{CM}	$\frac{I_{CM}}{E_{CM}}$
Erreur	Σ	u	E_{CM}	
Totale	$\underline{y}'\underline{y}$	n		

où :

n est le nombre total des observations
p est le nombre de paramètres dans (2)
m est le nombre de restrictions
u est le nombre de degrés de liberté de la somme de carrés relative à l'erreur et est donné par :

$$u = \sum_{\substack{i,j,k \\ \cap ijk \neq 0}} (n_{ijk} - 1)$$

Tests des interactions doubles

Pour tester l'absence d'interactions doubles nous utiliserons la méthode que l'on emploie lorsqu'on veut tester la nullité d'un ensemble de paramètres dans une régression.

Soit ω l'hypothèse de l'absence d'interactions doubles et soit

$$\underline{y} = X^*\underline{\beta}^* + \underline{e}^*$$

le modèle correspondant à cette hypothèse. Désignons par $\underline{b}^*$ l'estimation de $\underline{\beta}^*$ obtenue par la méthode des moindres carrés. Nous aurons alors l'analyse de variance suivante :

Source de variabilité	SC	DL	CM	F
$\underline{b}$	$\underline{b}'X'\underline{y}$	$p - m$		
$\underline{b}^*$	$\underline{b}^{*\prime}X^{*\prime}\underline{y}$	$p - m' - m$		
Interaction-2	$\underline{b}'X'\underline{y} - \underline{b}^{*\prime}X^{*\prime}\underline{y}$	m'	I_{CM}	$\frac{I_{CM}}{E_{CM}}$
Interaction-3	$\underline{y}'\underline{y} - \underline{b}'X'\underline{y} - \Sigma$	$n - p + m - u$		
Erreur	Σ	u	E_{CM}	
Total	$\underline{y}'\underline{y}$	n		

où m' est le nombre de degrés de liberté de l'interaction considérée et peut être calculée à l'aide de la formule suivante :

$$m' = (r' - 1)(r'' - 1)$$

r' et r'' étant les nombres de niveaux des facteurs considérés.

Test des effets principaux

Nous utiliserons exactement la même méthode que nous avons utilisée pour le test des interactions doubles ; nous supposerons que le test des interactions double et triple a été effectué et qu'il a donné un résultat négatif c'est-à-dire que l'hypothèse d'absence de ces interactions a été acceptée.

Soit ω° l'hypothèse d'absence des effets principaux et soit

$$\underline{y} = X^\circ\underline{\beta}^\circ + \underline{e}^\circ$$

le modèle correspondant à cette hypothèse ; nous aurons alors l'analyse de variance suivante :

Source de variabilité	SC	DL	CM	F
$\underline{b}$	$\underline{b}'X'\underline{y}$	$p' - m''$		
$\underline{b}^\circ$	$\underline{b}^{\circ\prime}X^{\circ\prime}\underline{y}$	$p' - m'' - m'''$		
Effet principal	$\underline{b}'X'\underline{y} - \underline{b}^{\circ\prime}X^{\circ\prime}\underline{y}$	m'''	EP_{CM}	$\frac{EP_{CM}}{E_{CM}}$
Reste	$\underline{y}'\underline{y} - \underline{b}'X'\underline{y}$	$n - p' - m'''$	E_{CM}	
Total	$\underline{y}'\underline{y}$	n		

où p' est le nombre de paramètres dans le modèle général sans interaction et peut être calculé à l'aide de la relation suivante :

$$p' = r + s + t.$$

m'' est le nombre de restrictions dans ce modèle et est égal au nombre de facteurs envisagés, soit trois dans notre cas.

m''' est le nombre de degrés de liberté du facteur considéré, soit $(r''' - 1)$ si r''' désigne le nombre de niveaux de ce facteur.

Remarquons encore que $\underline{b}$ désigne l'estimation de β dans le modèle où toutes les interactions sont supposées nulles. A titre de simplification nous avons utilisé la même notation que pour le modèle (2) donné au paragraphe 1.

4. – DESCRIPTION GENERALE DU PROGRAMME (1)

Remarquons tout d'abord que pour effectuer les diverses analyses de variance mentionnées précédemment, il faut être en possession des matrices que nous avons désignées par X,X*, X°, L', L*' et L°'. Afin d'éviter le travail long et fastidieux qui consiste à inscrire puis à perforer les éléments de ces diverses matrices, nous avons inclus dans le programme principal un sous-programme permettant :

1/ de calculer les matrices X et L' lorsqu'on connaît r, s et t et les effectifs des cellules.

2/ de calculer X*, X°, L*' et L°' à partir des matrices X et L'.

Suivant les dimensions de la matrice G définie par la relation (10) et la capacité de l'ordinateur utilisé, on utilisera l'une ou l'autre des méthodes indiquées. Bien qu'il eût été possible de réunir les deux méthodes dans un même programme, il nous a paru plus pratique d'avoir deux programmes distincts. Le langage utilisé est le FORTRAN IV G.

Le paquet de cartes des données est constitué de :

1/ Une carte de contrôle contenant les données générales du problème traité (nombre de niveaux, nombre total d'observations, dimensions de G)

2/ Des cartes contenant les valeurs des observations

3/ Sept cartes de contrôle permettant d'obtenir les six analyses de variance correspondant d'une part aux tests des trois interactions et d'autre part aux tests des trois effets principaux.

En modifiant légèrement le programme on peut facilement obtenir les valeurs des paramètres pour tout modèle supposant la nullité de l'un ou plusieurs des effets principaux ou des interactions.

Les résultats fournis par le programme sont :

1/ Les valeurs des paramètres pour le modèle complet donné par la relation (2)

2/ Les trois tableaux d'analyse de la variance pour le test des trois interactions doubles suivis des valeurs des paramètres dans le modèle correspondant ; dans chacun des trois tableaux apparaît le test de l'interaction triple.

3/ Les trois tableaux d'analyse de la variance pour les tests des trois effets principaux avec les valeurs des paramètres pour le modèle correspondant.

- - - - - - - - - - - - - - - - -

(1) Le programme avec une description détaillée peut être obtenu sur simple demande auprès de l'auteur à l'adresse suivante : Institut de statistique mathématique de l'Université de Genève, 2-4 rue du Lièvre, Case postale 124, 1 211 Genève 24 Suisse.

A titre d'exemple nous donnons les résultats obtenus en prenant les données du problème de Stevens.

En ce qui concerne les temps de calcul, il est évident qu'ils dépendent essentiellement de l'ordinateur utilisé ; d'autre part la méthode employant les inverses généralisées est beaucoup plus longue que la méthode employant les matrices régulières. Il est donc évident que si la capacité de l'ordinateur utilisé le permet on employera de préférence la seconde méthode.

A titre indicatif, les temps nécessaires pour traiter le probleme de Stevens, pour lequel les niveaux sont $r = 2$, $s = 4$ et $t = 4$, sont les suivants sur un ordinateur IBM 360/75 :

Première méthode : 22′ 30″
Deuxième méthode : 3′ 30″

BIBLIOGRAPHIE

[1] CHAKRAVARTI I.M. – *On generalized inverses in a linear associative algebra and their applications in the analysis of a class of designs,* 1971, Institute of Statistics Mimeo Series 766, University or North Carolina, Chapel Hill.

[2] GRAYBILL, F.A. – *An Introduction to Linear Statistical Models,* 1961, McGraw-Hill, New-York.

[3] GRAYBILL, F.A. – *Introduction to Matrices with Applications in Statistics,* 1969, Wadsworth, Belmont, California.

[4] GREVILLE, T.N.E. – Some applications of a matrix, 1960, *SIAM Review,* pp. 15-22.

[5] KEMPTHORNE, O. – *The Design and Analysis of Experiments,* 1952, John Wiley, New-York.

[6] SCHEFFE, H. – *The Analysis of Variance,* 1959, John Wiley, New-York.

[7] STEVENS, W.L. – Statistical analysis of a non-orthogonal tri-factorial experiment, 1948, *Biometrika,* 35, pp. 346-367.