

REVUE DE STATISTIQUE APPLIQUÉE

T. CALÍNSKY

M. LEJEUNE

Choix du nombre de fonctions linéaires discriminantes

Revue de statistique appliquée, tome 46, n° 1 (1998), p. 31-44

http://www.numdam.org/item?id=RSA_1998__46_1_31_0

© Société française de statistique, 1998, 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/>

CHOIX DU NOMBRE DE FONCTIONS LINÉAIRES DISCRIMINANTES

T. Calínsky (1), M. Lejeune (2)

(1) *Dept. of Mathematical and Statistical Methods, Agricultural University of Poznań, Poland*

(2) *CEDRIC, Conservatoire National des Arts et Métiers, Paris*

RÉSUMÉ

On s'intéresse à la détermination de la dimension de l'espace engendré par les fonctions linéaires discriminantes pour séparer k populations multivariées. Cette détermination est importante pour ne pas conserver de bruit dans les analyses effectuées dans le cadre de l'analyse discriminante. A cet effet plusieurs procédures relativement proches ont été proposées qui reposent sur des tests séquentiels d'hypothèses de dimensions croissantes. A chaque étape les valeurs critiques sont issues des distributions asymptotiques des statistiques utilisées. Nous montrons qu'en raison des approximations effectuées le risque nominal de surestimation de la dimension peut être sensiblement sous-évalué. On développe ici une procédure qui garantit le niveau nominal α vis-à-vis de ce danger de surestimation. Elle consiste en une transcription du problème en une série de tests d'hypothèses linéaires pour les vecteurs des moyennes des populations, dont les statistiques se trouvent être identiques à celles de la procédure usuelle introduite par Rao, mais ici leurs quantiles quasi exacts sont utilisés. La procédure proposée est comparée à celle de Rao au moyen de simulations. Elle est préférable surtout lorsque le nombre de dimensions est faible.

Mots-clés : *Analyse discriminante, fonctions linéaires discriminantes, hypothèses linéaires multivariées, trace de Lawley-Hotelling, hypothèses de dimensionnalité.*

ABSTRACT

The determination of the dimension of the space spanned by the linear discriminant functions for k multivariate populations is considered. This determination is crucial in order to avoid noise in the analyses linked to discriminant analysis. Several procedures of a similar nature have been proposed, which rest on sequential tests of hypotheses on increasing dimensions. At each step critical values are computed from the asymptotic chi-square distributions of the statistics. We show that, due to the approximations, the actual risk of overestimating the true dimension may exceed the nominal value quite significantly. In this paper a procedure is developed which ensures that the nominal α -level is not exceeded. This procedure is derived by rewriting the dimensionality hypotheses into linear hypotheses on the vectors of means of the populations, the statistics of which turn out to be identical to those of the classical approach due to Rao, but their quasi exact distributions are now used. Various simulation results are given and analyzed to compare the proposed procedure with Rao's procedure. The new procedure is quite preferable, especially when the number of dimensions is small.

Keywords : *Discriminant analysis, linear discriminant functions, multivariate linear hypothesis, Lawley-Hotelling trace, dimensionality hypotheses.*

1. Introduction

Le problème auquel nous nous intéressons ici est celui du choix du nombre de fonctions discriminantes significatives en analyse factorielle discriminante linéaire. Ce choix revêt une importance particulière dans la mesure où, au même titre que le choix des variables observables retenues, il influe sur le taux d'erreur de classement. En effet l'utilisation de composantes non significatives introduit du bruit dans la détermination de la procédure de classement d'individus nouveaux, nuisible pour la reproductibilité des résultats et, donc, susceptible de dégrader ce taux d'erreur. Déjà Fisher (1938) avait pressenti l'intérêt de ce problème en posant les premières bases d'un test simultané sur les valeurs propres inhérentes à l'analyse. Par la suite diverses procédures de tests ont été proposées pour ce problème, qui reposent cependant toutes sur des lois asymptotiques et donc approximatives. Il en résulte que le risque de première espèce pourra être sous-évalué et conduire à une surestimation du nombre utile de fonctions discriminantes.

Dans cet article nous développons une procédure dont l'avantage essentiel est qu'elle garantit le niveau nominal du test en recourant à des lois à tailles d'échantillons finies, permettant ainsi de contrôler le risque de surestimation du nombre de composantes. Cette procédure est une application d'un résultat établi par Caliński et Lejeune (1996) dans le cadre plus général de l'analyse de variance multivariée, pour déterminer la dimensionnalité de toute fonction paramétrique linéaire. Elle transcrit le problème en une suite d'hypothèses emboîtées, linéaires par rapport aux vecteurs des moyennes des k populations sous-jacentes, hypothèses pour lesquelles la statistique de la trace de Lawley-Hotelling a une distribution connue avec une très bonne approximation.

Notons que la dimension de l'espace des fonctions discriminantes est la dimension du sous-espace affine contenant les moyennes dans $\mathbb{R}^p$ des populations, où p est le nombre de variables. Le problème posé est donc également celui de l'analyse des relations entre les vecteurs de moyennes en analyse de variance multivariée à une voie, en utilisant des représentations en projection sur des dimensions en nombre pertinent. Cette approche a été décrite pour l'étude des effets des facteurs en analyse de variance multivariée par Caliński et Lejeune (1995), avec des applications dans le domaine de l'expérimentation agronomique.

Dans la section qui suit nous rappelons les notions de base de l'analyse discriminante linéaire puis, en section 3, nous présenterons les procédures classiques et, plus particulièrement, le test de dimensionnalité dû à Rao. La section 4 introduira la procédure que nous proposons. En section 5 nous exposerons la méthode de simulation utilisée pour comparer l'approche classique et l'approche proposée ici, et nous analyserons les résultats de ces simulations.

2. Rappels de l'analyse discriminante et notations

On considère un vecteur aléatoire dans $\mathbb{R}^p$ distribué selon des lois normales multivariées $\mathcal{N}_p(\mu_j, \Sigma)$, $j = 1, \dots, k$ sur k populations de même matrice des variances-covariances Σ , laquelle vérifie $\text{rang}(\Sigma) = p$. On dispose d'échantillons aléatoires de tailles respectives $n_1, n_2, \dots, n_k$ et on note $\mathbf{x}_{ij}$ le i -ème vecteur

d'observations de la j -ème population. On définit la $(p \times p)$ -matrice des sommes de carrés et produits croisés intra-échantillons :

$$\mathbf{W} = \sum_{j=1}^k \sum_{i=1}^{n_j} (\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ij} - \bar{\mathbf{x}}_j)',$$

et la $(p \times p)$ -matrice des sommes de carrés et produits croisés inter-échantillons :

$$\mathbf{B} = \sum_{j=1}^k n_j (\bar{\mathbf{x}}_j - \bar{\mathbf{x}})(\bar{\mathbf{x}}_j - \bar{\mathbf{x}})',$$

où $\bar{\mathbf{x}}_j = (1/n_j) \sum_{i=1}^{n_j} \mathbf{x}_{ij}$, $\bar{\mathbf{x}} = (1/n) \sum_{j=1}^k n_j \bar{\mathbf{x}}_j$ et $n = \sum_{j=1}^k n_j$. Les rangs de $\mathbf{W}$ et $\mathbf{B}$ sont respectivement p et $s = \min(p, k-1)$. Les fonctions linéaires discriminantes successives sont définies, pour tout $\mathbf{x} \in \mathbb{R}^p$, par $\mathbf{a}'_1 \mathbf{x}, \mathbf{a}'_2 \mathbf{x}, \dots, \mathbf{a}'_s \mathbf{x}$ où $\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_s$ sont les vecteurs propres normés de la matrice $\mathbf{W}^{-1} \mathbf{B}$ correspondant aux valeurs propres non nulles $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_s$. Pour la ℓ -ème fonction le pouvoir discriminant est mesuré par $\lambda_\ell = \mathbf{a}'_\ell \mathbf{B} \mathbf{a}_\ell / \mathbf{a}'_\ell \mathbf{W} \mathbf{a}_\ell$ ou par le rapport de corrélation $\lambda_\ell / (1 + \lambda_\ell)$.

Etant donné que ce sont les valeurs prises par ces fonctions discriminantes qui vont permettre de classer dans telle ou telle population un individu dont on connaît la valeur de $\mathbf{x}$, peut se poser le problème de savoir à partir de quelle valeur de ℓ la contribution des fonctions restantes n'est plus significative. Pour cela il va donc falloir tester des hypothèses sur des paramètres théoriques correspondants, caractérisant les populations.

Du point de vue théorique, les fonctions linéaires discriminantes sont données par les vecteurs propres normés $\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \dots, \boldsymbol{\nu}_s$ de la matrice de rang s : $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1} \boldsymbol{\Delta}$, où $\boldsymbol{\Delta}$ est la $(p \times p)$ -matrice de non centralité

$$\boldsymbol{\Delta} = \sum_{j=1}^k n_j (\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})(\boldsymbol{\mu}_j - \bar{\boldsymbol{\mu}})'$$

avec $\bar{\boldsymbol{\mu}} = \frac{1}{n} \sum_{j=1}^k n_j \boldsymbol{\mu}_j$. Les contributions relatives des fonctions linéaires discrimi-

nantes sont les valeurs propres de $\boldsymbol{\Omega}$: $\{\gamma_1, \gamma_2, \dots, \gamma_s\}$. La suite $\{\lambda_1, \lambda_2, \dots, \lambda_s\}$ est une statistique asymptotiquement exhaustive pour la suite des valeurs propres $\{\gamma_1, \gamma_2, \dots, \gamma_s\}$ (cf. Backhouse et McKay, 1982) et estime cette dernière, moyennant une constante multiplicative $(n-k)$. De même $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_s\}$ estime la suite $\{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \dots, \boldsymbol{\nu}_s\}$ des vecteurs propres normés de $\boldsymbol{\Omega}$ [voir par exemple Krishnaiah et Kanal (1982, p. 883-892) ou Tomassone *et al.* (1988, p. 52-55)].

Notons que, par commodité, l'estimation est ici effectuée de façon conditionnelle aux n_j , considérés comme fixés. Pour une approche non conditionnelle on

remplace Δ par la matrice $\mathbf{M}$ vers laquelle converge $\frac{1}{n} \Delta$ presque sûrement, les probabilités *a priori* d'appartenance aux populations, p_j , se substituant aux n_j . Alors les λ_j et $\mathbf{a}_j$ convergent presque sûrement vers les valeurs et vecteurs propres de la matrice $\Sigma^{-1}\mathbf{M}$.

Déterminer le nombre de fonctions linéaires discriminantes pertinent revient donc à déterminer le nombre de valeurs propres γ_j non nulles. De façon équivalente il s'agit de déterminer le rang de $\Omega = \Sigma^{-1}\Delta$ et donc de Δ , lequel donne la dimension du sous-espace affine de $\mathbb{R}^p$ contenant $\mu_1, \mu_2, \dots, \mu_k$. En particulier, l'égalité des vecteurs des moyennes correspond au rang 0 et à des valeurs propres toutes nulles.

3. Le test de dimensionnalité de Rao

En introduction nous avons mentionné qu'il existe plusieurs tests classiques pour notre problème. Les statistiques de ces tests sont, en fait, toutes dérivées de façon identique de celles utilisées pour tester l'égalité des vecteurs des moyennes $H_0 : \mu_1 = \mu_2 = \dots = \mu_k$ ou encore $H_0 : \Omega = 0$. Ces statistiques reposent sur les valeurs propres de la matrice $\mathbf{W}^{-1}\mathbf{B}$. Ce sont, pour l'essentiel, la statistique de Roy

$c_{max} = \lambda_1$, la trace de Lawley-Hotelling $T_0^2 = (n - k) \sum_{i=1}^s \lambda_i$, le rapport de Wilks

$U = \prod_{i=1}^s 1/(1 + \lambda_i)$ et la trace de Pillai $V = \sum_{i=1}^s \lambda_i/(1 + \lambda_i)$.

Nous nous intéresserons ici uniquement au test fondé sur la trace de Lawley-Hotelling qui constitue le test de Rao. En effet l'étude de Backhouse et McKay (1982) a montré que les tests correspondant à ces différentes statistiques donnent des résultats très semblables. On trouvera, par exemple, dans Tomassone *et al.* (1988, p.156-157) les développements relatifs à la statistique de Wilks.

Rao (1973, section 8c.6) propose de tester séquentiellement les s hypothèses suivantes :

$$\begin{aligned} H_0 & : \gamma_1 = \gamma_2 = \dots = \gamma_s = 0, \\ H_1 & : \gamma_1 > \gamma_2 = \dots = \gamma_s = 0, \\ & \dots \\ H_d & : \gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_d > \gamma_{d+1} = \dots = \gamma_s = 0, \\ & \dots \\ H_{s-1} & : \gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_{s-1} > \gamma_s = 0. \end{aligned}$$

Pour tester H_d , la statistique utilisée est déduite de la statistique du rapport de vraisemblance établie pour Σ connue et dont la distribution asymptotique, selon la théorie du rapport de vraisemblance, est, sous H_d , une loi du Khi-deux à $(p - d)(k - d - 1)$ degrés de liberté. Comme Σ est, en pratique, inconnue on lui substitue son estimateur $\frac{1}{n-k} \mathbf{W}$ pour aboutir à la statistique :

$$T_{0,d}^2 = (n - k) \sum_{i=d+1}^s \lambda_i.$$

On introduit ainsi un deuxième niveau d'approximation qui donne une variabilité plus élevée à la statistique utilisée. On peut donc s'attendre à ce que le niveau α choisi nominalement soit en réalité optimiste (test dit «anticonservateur»). Pour les autres types de tests les statistiques correspondantes sont :

$$c_{max,d} = \lambda_{d+1}, \quad U_d = \prod_{i=d+1}^s 1/(1 + \lambda_i) \quad \text{et} \quad V_d = \sum_{i=d+1}^s \lambda_i/(1 + \lambda_i).$$

La procédure consiste à examiner, successivement et à chaque étape au niveau α , la significativité des statistiques $T_0^2, T_{0,1}^2$, etc., jusqu'au moment où $T_{0,d}^2$ devient non significative, ce qui conduit à choisir le rang d .

En fait, on teste chaque hypothèse de façon unilatérale : on rejette H_d , par exemple, si $T_{0,d}^2$ est supérieur au quantile $1 - \alpha$ de la loi du khi-deux appropriée. Cela implique que l'alternative est donc : $\text{rang}(\Omega) > d$, et non pas : $\text{rang}(\Omega) \neq d$. Or une telle procédure de tests multiples où les alternatives successives sont emboîtées est nécessairement conservatrice par rapport au risque de première espèce α utilisé à chaque étape (cf. Marcus *et al.*, 1976). On a donc la garantie que la probabilité de rejeter l'hypothèse de rang d , si elle est vraie, au profit d'un rang *supérieur*, ne dépassera pas α . En d'autres termes le risque d'une surévaluation du nombre de valeurs propres non nulles est contrôlé. Toutefois ceci n'est valable que si, à chaque étape, le risque α est effectif. Or les tests classiques utilisant des approximations asymptotiques la garantie du risque nominal α n'existe pas. En fait, on peut voir sur le développement de ces approximations donné par Anderson (1984, section 8.6.2) que le quantile exact est celui du Khi-deux préconisé plus un terme positif et un terme en $O((n - k)^{-2})$. Ces tests seront donc généralement anticonservateurs, ce que montre explicitement la table D15 de Seber (1984).

La procédure que nous proposons a pour objet de garantir le risque α à chaque étape et donc de contrôler le risque α global.

4. La procédure de test proposée

Le problème est donc de décider du rang de $\Sigma^{-1}\Delta$, c'est-à-dire du rang de $\Delta = \sum_{j=1}^k n_j(\mu_j - \bar{\mu})(\mu_j - \bar{\mu})'$. Introduisons la $(k \times p)$ -matrice paramétrique $\Xi = (\mu_1, \mu_2, \dots, \mu_k)'$ et la $(k \times k)$ -matrice de centrage $C = I_k - \frac{1}{n}\mathbf{1}_k\mathbf{n}'$, avec $\mathbf{1}_k = (1, 1, \dots, 1)'$ et $\mathbf{n}' = (n_1, n_2, \dots, n_k)$, d'où $C\Xi = (\mu_1 - \bar{\mu}, \mu_2 - \bar{\mu}, \dots, \mu_k - \bar{\mu})'$. Comme $\Delta = (C\Xi)' \text{diag}(\mathbf{n}) C\Xi$, le rang de Δ est égal au rang de $C\Xi$, ce qui transcrit simplement le fait que le rang de Δ donne la dimension de l'hyperplan contenant les moyennes des populations dans $\mathbb{R}^p$.

Nous allons fonder notre décision sur le résultat du test séquentiel des hypothèses :

$$\begin{aligned}
 H_0 & : \text{rang}(\Delta) = 0 \\
 H_1 & : \text{rang}(\Delta) \leq 1 \\
 & \dots \\
 H_d & : \text{rang}(\Delta) \leq d \\
 & \dots \\
 H_{s-1} & : \text{rang}(\Delta) \leq s-1.
 \end{aligned}$$

Si les hypothèses $H_0, H_1, \dots, H_{d-1}$ sont rejetées l'une après l'autre au risque α et si H_d est la première hypothèse à ne pas être rejetée au même niveau de risque, alors on conclura que la dimension est d . Du fait de l'emboîtement de ces hypothèses cette procédure séquentielle est une procédure de test dont le risque de première espèce est égal ou inférieur à α (cf. section 3).

L'hypothèse H_d peut s'exprimer de la façon suivante :

il existe une $(p \times q)$ -matrice $\mathbf{Z}_d$ de rang plein $q \leq d$ telle que pour toute $(p \times (p-q))$ -matrice $\mathbf{A}_d$ de rang $p-q$ orthogonale à $\mathbf{Z}_d$, c'est-à-dire vérifiant $\mathbf{Z}_d' \mathbf{A}_d = \mathbf{0}$, on ait $\mathbf{C} \Xi \mathbf{A}_d = \mathbf{0}$.

Notons que les colonnes de $\mathbf{Z}_d$ engendrent le sous-espace de $\mathbb{R}^p$ contenant les colonnes de $(\mathbf{C} \Xi)' = (\mu_1 - \bar{\mu}, \mu_2 - \bar{\mu}, \dots, \mu_k - \bar{\mu})$.

Pour une matrice $\mathbf{A}_d$ donnée l'équation $\mathbf{C} \Xi \mathbf{A}_d = \mathbf{0}$ est une hypothèse linéaire sur la matrice paramétrique Ξ du modèle. Dans le contexte général du modèle linéaire gaussien cette hypothèse peut être testée par la statistique de la trace de Lawley-Hotelling :

$$T_o^2(\mathbf{A}_d) = (n-k) \text{trace}[(\mathbf{A}_d' \mathbf{W} \mathbf{A}_d)^{-1} \mathbf{A}_d' \mathbf{B} \mathbf{A}_d].$$

Sous l'hypothèse nulle $\mathbf{C} \Xi \mathbf{A}_d = \mathbf{0}$, la loi de cette statistique ne dépend que de k, p et q , et nous noterons $T_{o,1-\alpha}(k, p, q)$ son quantile $1 - \alpha$. Un des intérêts de cette statistique est que cette loi peut être très précisément approchée, selon un résultat établi par McKeon(1974), au moyen d'une loi de Fisher-Snedecor que nous expliciterons en section 5. Pour apprécier la précision de cette approximation voir Seber (1984, p. 39).

A supposer, pour l'heure, que $\mathbf{Z}_d$ soit connue, on acceptera H_d au risque α si :

$$\sup_{\mathbf{Z}_d' \mathbf{A}_d = \mathbf{0}} (n-k) \text{trace}[(\mathbf{A}_d' \mathbf{W} \mathbf{A}_d)^{-1} \mathbf{A}_d' \mathbf{B} \mathbf{A}_d] < T_{o,1-\alpha}(k, p, q).$$

Comme nous testons $H_0, H_1, \dots$ séquentiellement il suffit, pour tester H_d , de considérer $q = d$ pour définir les matrices $\mathbf{Z}_d$ et $\mathbf{A}_d$, et de choisir le niveau critique $T_{o,1-\alpha}(k, p, d)$. Toutefois, étant donné que $\mathbf{Z}_d$ est inconnue, nous devons nous contenter d'une minoration de cette statistique et l'on opérera au risque inférieur ou égal à α en adoptant la règle de rejet :

$$\inf_{\mathbf{Z}_d} \sup_{\mathbf{Z}_d' \mathbf{A}_d = \mathbf{0}} (n-k) \text{trace}[(\mathbf{A}_d' \mathbf{W} \mathbf{A}_d)^{-1} \mathbf{A}_d' \mathbf{B} \mathbf{A}_d] > T_{o,1-\alpha}(k, p, d),$$

en prenant l'infimum sur l'ensemble des $(p \times d)$ -matrices de rang d . La statistique ci-dessus étant, en fait, égale à $(n - k) \sum_{j=d+1}^s \lambda_j$ (cf. Caliński et Lejeune, 1996, page 6), on a finalement :

$$\text{Prob}((n - k) \sum_{j=d+1}^s \lambda_j \geq T_{o,1-\alpha}(k, p, d) \mid H_d) \leq \alpha ,$$

ce qui détermine la règle de rejet de H_d . Notons que le fait que la statistique utilisée ici soit majorée, avec probabilité 1, par la variable aléatoire caractérisée par les quantiles $T_{o,1-\alpha}(k, p, d)$, a été établi par une autre voie par Schott (1984).

On retombe donc, par une voie bien différente, sur la statistique de test utilisée par Rao. Toutefois la procédure diffère pour ce qui concerne la détermination des valeurs critiques. Dans le cas classique on recourt à la loi asymptotique comme approximation alors qu'ici on utilise la loi exacte (ou quasi exacte) pour s'assurer d'avoir, dès l'hypothèse de rang nul H_0 , un risque de première espèce parfaitement contrôlé.

Nous sommes ici en présence d'une procédure qui est conservatrice du fait qu'elle minore, à chaque étape, la statistique de test, alors que les tests classiques opèrent par des approximations de nature anticonservatrice. Caliński et Lejeune (1996) montrent, en utilisant une variante de démonstration, que la perte de puissance due à la minoration s'estompe quand la taille de l'échantillon global augmente. Par ailleurs rappelons que pour toute procédure séquentielle de tests d'hypothèses emboîtées, le niveau α global se réduit au fur et à mesure que l'on avance dans les étapes et donc, ici, dans les dimensions croissantes.

Les simulations qui suivent vont nous permettre :

- tout d'abord de vérifier que notre procédure garantit bien le niveau α , c'est-à-dire que l'approximation de McKeon est effectivement extrêmement précise,
- de mesurer le degré de conservatisme de cette procédure, c'est-à-dire le prix payé pour garantir α , en terme de perte de puissance, mesurée par la propension à sous-estimer la vraie dimension,
- d'évaluer l'effet de l'approximation d'une procédure classique - celle de Rao - et de l'incidence de son caractère anticonservateur,
- de comparer les deux procédures dans différentes situations.

Nous nous sommes limités à la comparaison avec la procédure de Rao, parmi les procédures classiques. En effet, s'il est vrai qu'aucune des procédures classiques ne domine absolument les autres, les simulations effectuées antérieurement par Backhouse et McKay (1982) pour les comparer ont montré des variations mineures entre elles et il s'est avéré que, globalement, la procédure de Rao était sans doute la plus intéressante. Nous nous référerons d'ailleurs abondamment à cet article dans nos propres simulations.

5. Comparaisons par simulations

Bien que le nombre de situations intéressantes soit apparemment multiple, il est possible, sans perte de généralité et en suivant Backhouse et McKay, de se limiter aux cas où $\Sigma = \mathbf{I}$ et où Δ est une matrice diagonale contenant directement les valeurs propres d'intérêt $\gamma_1, \gamma_2, \dots, \gamma_s$ (voir également à ce propos Seo *et al.*, 1995).

Sachant que la matrice $\mathbf{W}$ suit une loi de Wishart de paramètres $\mathbf{I}_p$ et $n - k$, une réalisation de $\mathbf{W}$ peut être simulée par le produit $\mathbf{A}\mathbf{A}'$ où $\mathbf{A}$ est une $(p \times (n - k))$ -matrice de réalisations indépendantes issues de la loi $\mathcal{N}(0, 1)$. La matrice $\mathbf{B}$ est indépendante de $\mathbf{W}$ et suit une loi de Wishart de paramètres $\mathbf{I}_p$ et $k - 1$ avec matrice de non centralité Δ . Elle peut donc être simulée par le produit $(\mathbf{A} + \mathbf{D})(\mathbf{A} + \mathbf{D})'$ où $\mathbf{D}$ est une $(p \times (k - 1))$ -matrice de réalisations indépendantes de la loi $\mathcal{N}(0, 1)$ et $\mathbf{A}$ est la $(p \times (k - 1))$ -matrice diagonale telle que $\Delta = \mathbf{A}\mathbf{A}'$, c'est-à-dire dont les valeurs diagonales sont $\gamma_1^{1/2}, \gamma_2^{1/2}, \dots, \gamma_s^{1/2}$.

Dans les deux procédures comparées, celle de Rao et celle que nous proposons, la suite de statistiques utilisées est

$$\left\{ T_{0,d}^2 = (n - k) \sum_{j=d+1}^s \lambda_j, \quad d = 0, 1, \dots, s - 1 \right\}$$

où les λ_j sont les valeurs propres de la matrice $\mathbf{W}^{-1}\mathbf{B}$. Pour la procédure de Rao le niveau critique, à l'étape du test de la dimension d , est un quantile de la loi du Khi-deux à $(p - d)(k - d - 1)$ degrés de liberté, alors que pour la nouvelle procédure on utilise, selon l'approximation de McKeon, un quantile de la loi de Fisher-Snedecor à a et b degrés de liberté pour $T_{0,d}^2/c$ où :

$$\begin{aligned} a &= (p - d)(k - 1) \\ b &= 4 + \frac{a + 2}{f - 1} \quad \text{et} \quad f = \frac{(n - p + d - 2)(n - k - 1)}{(n - k - p + d - 3)(n - k - p + d)} \\ c &= \frac{a(b - 2)}{b(n - k - p + d - 1)}. \end{aligned}$$

Nous avons repris les situations considérées par Backhouse et McKay en les complétant avec des situations de rang 0 ou 1, absentes de leur étude. Les valeurs propres théoriques choisies se situent entre 0 et 100. Sachant que les valeurs propres empiriques sont *grosso modo* de l'ordre des valeurs propres théoriques divisées par $n - k$, on voit, à titre indicatif, qu'à une valeur théorique de 50 correspond une valeur empirique de 1 pour $n = 50$, et donc un rapport de corrélation de 50 % (ou pour une valeur propre de 20 : 29 %). La taille d'échantillon de référence est 50 mais la taille 500 a été également introduite pour apprécier les évolutions. Pour chaque situation 1000 répétitions ont été opérées.

Les résultats sont donnés dans les tableaux 1 et 2 sous la forme de fréquences relatives, parmi les 1 000 répétitions, des choix de dimension auxquels a conduit

la procédure. La fréquence figurant en gras correspond à celle du bon choix. Le cumul des fréquences à droite de celle-ci représente la fréquence de surestimation et, donc, est une estimation du niveau α réel. Pour les situations identiques à celles de Backhouse et McKay nous retrouvons bien les mêmes résultats.

Le premier constat est que l'excellente précision de l'approximation de MacKeon est confirmée puisque le niveau α est toujours vérifié dans notre approche, à quelques faibles variations près imputables au nombre limité de répétitions. En revanche la procédure classique peut donner des niveaux assez nettement au-delà du niveau nominal lorsqu'on est en présence d'une dimension réduite. Par exemple :

- pour le cas 2 avec rang nul, $\alpha = 0.10$ nominal donne en fait 0.21
- pour le cas 4 avec rang nul, $\alpha = 0.25$ nominal donne en fait 0.38
- pour le cas 4 avec rang 1 et valeur propre 50, $\alpha = 0.25$ nominal donne en fait 0.32.

Le caractère conservateur dû à l'emboîtement des hypothèses est très évident pour notre procédure : le risque de première espèce (qui est exact pour H_0 où la minoration n'intervient pas) décroît rapidement avec le nombre de tests effectués en séquence, c'est-à-dire avec la dimension. En comparant les cas 2 et 3 où n passe de 50 à 500 on peut apprécier grossièrement les influences respectives de la minoration et de l'emboîtement, sachant que l'effet de minoration s'estompe lorsque la taille de l'échantillon global s'accroît. On constate que le risque de première espèce varie peu entre ces deux situations et l'on est porté à croire que l'effet d'emboîtement domine celui de minoration.

Le prix à payer pour conserver α est une tendance à sous-estimer le nombre de composantes. Cette tendance n'est pas trop sensible quand le nombre de dimensions est faible, disons inférieur ou égal à 3, et lorsque la plus petite valeur propre reste importante, disons pas moins de 20 et plutôt 50. Par exemple, pour le cas 2 :

- avec valeurs propres 100/50/0/0, $\alpha = 0.1$, la dimension 2 est détectée avec probabilité (estimée) 0.98,
- avec valeurs propres 100/50/20/0, $\alpha = 0.1$, la dimension 3 est détectée avec probabilité 0.75,
- avec valeurs propres 50/20/10/0, $\alpha = 0.1$, la dimension 3 est détectée avec probabilité 0.21.

Il est normal que la tendance à sous-estimer soit plus prononcée lorsque les valeurs propres non nulles se rapprochent de 0, s'agissant d'une procédure par construction conservatrice, c'est-à-dire tendant à accepter plus que nécessaire.

D'un point de vue pratique il est intéressant de noter que l'efficacité de la procédure peut être accrue en élevant le niveau α . Ainsi pour $\alpha = 0.25$ les trois probabilités ci-dessus deviennent (*cf.* cas 4) : 0.94, 0.88 et 0.39.

Pour la procédure de Rao la diminution du niveau réel due à l'accroissement du nombre de dimensions apparaît également mais de façon moins rapide. La raison en est que le niveau réel de la procédure de Rao bénéficie en fait d'un effet heureux de compensation entre l'approximation asymptotique, qui le majore, et l'emboîtement des hypothèses, qui le minore. Ceci explique que la tendance à sous-estimer le nombre

TABLEAU 1

Fréquences relatives des choix de dimension pour 1 000 simulations

1^{er} ensemble de colonnes : valeurs propres théoriques2^e ensemble de colonnes : fréquence relative

du choix de la dimension 0, 1, 2, 3, 4 selon Rao

3^e ensemble de colonnes : fréquence relative du choix

de la dimension 0, 1, 2, 3, 4 selon la nouvelle procédure.

(le bon choix est indiqué **en gras**)

Cas 1 : n = 50 p = 4 k = 6 alpha = 0.05													
0	0	0	0	0.87	0.12	0.01	0.00	0.00	0.96	0.04	0.00	0.00	0.00
100	0	0	0	0.00	0.91	0.09	0.00	0.00	0.00	0.99	0.01	0.00	0.00
50	0	0	0	0.00	0.91	0.09	0.00	0.00	0.01	0.98	0.02	0.00	0.00
20	0	0	0	0.17	0.76	0.07	0.00	0.00	0.33	0.66	0.01	0.00	0.00
100	50	0	0	0.00	0.00	0.93	0.06	0.00	0.00	0.01	0.98	0.01	0.00
50	20	0	0	0.00	0.17	0.78	0.05	0.00	0.00	0.43	0.57	0.00	0.00
20	10	0	0	0.04	0.52	0.41	0.03	0.00	0.11	0.75	0.14	0.00	0.00
100	50	20	0	0.00	0.00	0.08	0.86	0.06	0.00	0.00	0.37	0.63	0.00
50	20	10	0	0.00	0.03	0.49	0.45	0.03	0.00	0.15	0.75	0.10	0.00
20	10	4	0	0.02	0.41	0.47	0.10	0.00	0.07	0.69	0.24	0.01	0.00
Cas 2 : n = 50 p = 4 k = 6 alpha = 0.10													
0	0	0	0	0.79	0.20	0.01	0.00	0.00	0.89	0.11	0.00	0.00	0.00
100	0	0	0	0.00	0.82	0.17	0.01	0.00	0.00	0.97	0.03	0.00	0.00
50	0	0	0	0.00	0.85	0.14	0.01	0.00	0.01	0.96	0.04	0.00	0.00
20	0	0	0	0.12	0.76	0.11	0.01	0.00	0.21	0.77	0.02	0.00	0.00
100	50	0	0	0.00	0.00	0.86	0.13	0.01	0.00	0.00	0.98	0.01	0.00
50	20	0	0	0.00	0.09	0.81	0.09	0.01	0.00	0.29	0.70	0.01	0.00
20	10	0	0	0.03	0.43	0.48	0.06	0.00	0.07	0.71	0.22	0.00	0.00
100	50	20	0	0.00	0.00	0.05	0.88	0.07	0.00	0.00	0.24	0.75	0.01
50	20	10	0	0.00	0.01	0.35	0.58	0.06	0.00	0.08	0.71	0.21	0.00
20	10	4	0	0.01	0.28	0.54	0.16	0.01	0.02	0.60	0.36	0.02	0.00
Cas 3 : n = 500 p = 4 k = 6 alpha = 0.10													
0	0	0	0	0.89	0.10	0.01	0.00	0.00	0.90	0.10	0.00	0.00	0.00
100	0	0	0	0.00	0.92	0.08	0.00	0.00	0.00	0.98	0.02	0.00	0.00
50	0	0	0	0.00	0.91	0.09	0.00	0.00	0.00	0.97	0.03	0.00	0.00
20	0	0	0	0.13	0.79	0.08	0.00	0.00	0.14	0.84	0.02	0.00	0.00
100	50	0	0	0.00	0.00	0.92	0.07	0.01	0.00	0.00	0.99	0.01	0.00
50	20	0	0	0.00	0.10	0.82	0.08	0.00	0.00	0.22	0.77	0.01	0.00
20	10	0	0	0.03	0.49	0.44	0.04	0.00	0.04	0.70	0.26	0.00	0.00
100	50	20	0	0.00	0.00	0.05	0.88	0.07	0.00	0.00	0.18	0.82	0.00
50	20	10	0	0.00	0.02	0.35	0.58	0.06	0.00	0.05	0.70	0.25	0.00
20	10	4	0	0.01	0.30	0.54	0.13	0.02	0.01	0.52	0.45	0.02	0.00

TABLEAU 2

Fréquences relatives des choix de dimension pour 1 000 simulations

1^{er} ensemble de colonnes : valeurs propres théoriques2^e ensemble de colonnes : fréquence relative

du choix de la dimension 0, 1, 2, 3, 4 selon Rao

3^e ensemble de colonnes : fréquence relative du choix

de la dimension 0, 1, 2, 3, 4 selon la nouvelle procédure.

(le bon choix est indiqué **en gras**)

Cas 4 : n = 50 p = 4 k = 6 alpha = 0.25													
0	0	0	0	0.62	0.32	0.05	0.01	0.00	0.76	0.24	0.00	0.00	0.00
100	0	0	0	0.00	0.67	0.28	0.04	0.01	0.00	0.89	0.11	0.00	0.00
50	0	0	0	0.00	0.68	0.26	0.05	0.01	0.00	0.88	0.12	0.00	0.00
20	0	0	0	0.14	0.70	0.22	0.04	0.01	0.08	0.84	0.08	0.00	0.00
100	50	0	0	0.00	0.00	0.71	0.23	0.06	0.00	0.00	0.94	0.06	0.00
50	20	0	0	0.00	0.04	0.75	0.17	0.04	0.00	0.15	0.82	0.03	0.00
20	10	0	0	0.00	0.28	0.55	0.13	0.04	0.02	0.56	0.41	0.02	0.00
100	50	20	0	0.00	0.00	0.01	0.76	0.23	0.00	0.00	0.10	0.88	0.02
50	20	10	0	0.00	0.00	0.18	0.64	0.18	0.00	0.02	0.57	0.39	0.02
20	10	4	0	0.00	0.13	0.49	0.30	0.08	0.01	0.35	0.56	0.07	0.01
Cas 5 : n = 50 p = 4 k = 10 alpha = 0.10													
0	0	0	0	0.73	0.25	0.02	0.00	0.00	0.91	0.09	0.00	0.00	0.00
100	0	0	0	0.00	0.81	0.18	0.01	0.00	0.00	0.94	0.06	0.00	0.00
50	0	0	0	0.00	0.82	0.17	0.01	0.00	0.02	0.93	0.05	0.00	0.00
20	0	0	0	0.15	0.72	0.12	0.01	0.00	0.34	0.64	0.02	0.00	0.00
100	50	0	0	0.00	0.01	0.86	0.13	0.01	0.00	0.03	0.95	0.02	0.00
50	20	0	0	0.00	0.19	0.73	0.08	0.01	0.00	0.48	0.50	0.02	0.00
20	10	0	0	0.04	0.54	0.38	0.04	0.00	0.16	0.70	0.13	0.00	0.00
100	50	20	0	0.00	0.00	0.14	0.79	0.08	0.00	0.00	0.43	0.56	0.01
50	20	10	0	0.00	0.05	0.54	0.37	0.04	0.00	0.21	0.67	0.12	0.00
20	10	4	0	0.03	0.41	0.47	0.08	0.01	0.12	0.66	0.22	0.01	0.00
Cas 6 : n = 50 p = 8 k = 5 alpha = 0.10													
0	0	0	0	0.64	0.32	0.04	0.00	0.00	0.92	0.08	0.00	0.00	0.00
100	0	0	0	0.00	0.73	0.25	0.02	0.00	0.00	0.99	0.01	0.00	0.00
50	0	0	0	0.00	0.74	0.24	0.02	0.00	0.02	0.97	0.01	0.00	0.00
20	0	0	0	0.07	0.71	0.21	0.01	0.00	0.32	0.68	0.01	0.00	0.00
100	50	0	0	0.00	0.00	0.82	0.16	0.02	0.00	0.04	0.96	0.00	0.00
50	20	0	0	0.00	0.11	0.75	0.13	0.01	0.00	0.60	0.40	0.00	0.00
20	10	0	0	0.03	0.41	0.50	0.06	0.01	0.14	0.80	0.06	0.00	0.00
100	50	20	0	0.00	0.00	0.08	0.79	0.13	0.00	0.00	0.76	0.24	0.00
50	20	10	0	0.00	0.02	0.44	0.48	0.06	0.00	0.31	0.68	0.01	0.00
20	10	4	0	0.01	0.28	0.54	0.15	0.02	0.08	0.78	0.14	0.00	0.00

de dimensions soit souvent moindre lorsque le nombre de dimensions dépasse 2 ou 3 et/ou que la dernière valeur propre est faible, de l'ordre de 20. Quoi qu'il en soit l'effet d'approximation reste dominé par l'effet conservateur inhérent à l'approche séquentielle. D'ailleurs le fait que l'approximation par la loi du Khi-deux opère systématiquement à un niveau supérieur au niveau nominal n'a pas été constaté par Backhouse et McKay parce que leur étude ne prend pas en considération, du moins dans les résultats publiés, des situations à 0 ou 1 valeur propre non nulle. C'est ce qui les amène à affirmer, à tort, que les tests opèrent sous le niveau nominal.

Pour ce qui concerne les évolutions de la procédure classique lorsque la taille de l'échantillon global s'accroît de 50 à 500, on constate que le risque α diminue nettement uniquement pour les valeurs propres élevées, disons toutes supérieures à 50 (voir cas 2 et 3). Ceci est à rapprocher du résultat établi par Bartlett (1947) : la loi limite des $\lambda_{d+1}, \dots, \lambda_s$ conditionnellement aux $\lambda_1, \dots, \lambda_d$ sous H_d , s'entend non seulement pour $n \rightarrow \infty$ mais aussi pour les $\lambda_1, \dots, \lambda_d$ tendant vers l'infini. Or, pour des γ_i données, les λ_i intègrent un facteur en $O_p(1/n)$ qui vient freiner la vitesse de convergence. On peut croire qu'il en va de même pour la loi limite du rapport de vraisemblance. Ce phénomène a été mis en évidence par Backhouse et McKay pour les différentes procédures classiques.

La juxtaposition des cas 2 et 6 permet de voir l'incidence de p , le nombre de variables. Pour la procédure de Rao l'accroissement de p de 4 à 8 augmente assez nettement le risque α . Par exemple, pour la situation de rang 0, il passe de 0.21 à 0.36 alors que le niveau nominal est de 0.10. Dans les deux procédures le risque de sous-estimation du nombre de composantes s'accroît.

La juxtaposition des cas 2 et 5 permet de voir l'incidence de k , le nombre de groupes. L'augmentation de ce nombre de 4 à 10 a également pour effet d'accroître le niveau α pour la procédure de Rao, mais cela est moins sensible que précédemment : toujours pour la situation de rang 0, il passe de 0.21 à 0.27. Ici aussi le risque de sous-estimation est accru partout.

6. Conclusion

La procédure que nous avons introduite a pour qualité essentielle de garantir le risque de surestimation du nombre réel de dimensions. Cette qualité n'est pas possédée, loin s'en faut parfois, par la procédure de Rao, ni par les autres procédures classiques dont les comportements restent très proches de cette dernière comme l'ont montré Backhouse et McKay (1982).

Comme pour toute procédure de tests multiples le niveau de conservatisme s'accroît avec le nombre de tests effectués, c'est-à-dire, ici, avec la dimension réelle. Il en découle une certaine propension à la sous-estimation de cette dimension et ceci d'autant plus que les dernières valeurs propres sont faibles, disons inférieures à 20.

Pour la procédure classique l'approximation asymptotique est anticonservatrice et vient donc atténuer l'effet conservateur des tests séquentiels. Il en résulte que pour un nombre de dimensions au-delà de 2 et des dernières valeurs propres petites, cette approche donne un risque de sous-estimation moins élevé que la procédure que nous

proposons. Il faut cependant remarquer que ce jeu de compensation est imprévisible et qu'il n'est pas possible d'en dégager des règles pratiques systématiques.

Pour réduire la sous-estimation de la dimension on peut envisager, pour la procédure proposée, de repartir d'un risque nominal plus élevé – par exemple 0.25 – si l'on s'attend à un nombre de dimensions au-delà de 2. L'idéal serait d'ailleurs de tirer profit de la décroissance rapide du risque en augmentant sa valeur nominale au fur et à mesure que l'on avance dans les rangs supérieurs. Ceci mériterait un examen plus approfondi. Alors que le principe de cette approche à risque évolutif est envisageable pour notre procédure grâce au contrôle du risque α à chaque étape, il n'en va pas de même pour les procédures classiques en raison des effets compensatoires mentionnés plus haut, qui confèrent un certain arbitraire à la valeur du niveau réel. Une autre voie d'approfondissement consisterait à s'attaquer directement au problème de la sous-estimation en considérant les hypothèses alternatives *ad hoc*.

Quoi qu'il en soit, la procédure que nous avons présentée est indéniablement très efficace pour les faibles dimensions, là-même où les procédures classiques recèlent un danger certain de surestimation. Nous y voyons encore un avantage dans la mesure où elle est généralisable à de nombreuses situations : test de dimensionnalité sur les effets d'un seul facteur dans l'analyse de variance multivariée à plusieurs voies, test de dimensionnalité d'effets d'interaction et toutes circonstances où l'on peut se ramener à une modélisation linéaire.

Références bibliographiques

- ANDERSON T.W. (1984), *An Introduction to Multivariate Statistical Analysis*, Wiley.
- BACKHOUSE A.R., McKAY R.J. (1982), Tests of dimensionality in multivariate analysis of variance. *Communications in Statistics : Theory and Methods*, **11**, 1003-1027.
- BARTLETT M.S. (1947), Multivariate Analysis, *JRSS, Suppl.* **9 Series B**, 176-197.
- CALIŃSKI T., LEJEUNE M. (1995), Factor analysis of matrices with applications to multivariate analysis of variance. *Série des Documents de Travail du Centre de Recherche en Economie et Statistique*, No **9517**. INSEE, CREST, Bureau 2105, 15 bd Gabriel Péri, 92245 Malakoff-France, 31 p.
- CALIŃSKI T., LEJEUNE M. (1996), A fresh look at testing hypotheses on dimensionality in the Manova model. *Série des Documents de Travail du Centre de Recherche en Economie et Statistique*, No **9632**. INSEE, CREST, Bureau 2105, 15 bd Gabriel Péri, 92245 Malakoff-France, 16 p.
- FISHER R.A. (1938), The statistical utilization of multiple measurements. *Annals of Eugenics*, **8**, 376-386.
- KRISHNAIAH P.R., KANAL L.N. eds. (1982), *Handbook of Statistics*, vol. 2, North Holland.
- MARCUS R., PERITZ E. and GABRIEL K.R. (1976), On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655-660.

- MCKEON J.J. (1974), F approximations to the distribution of Hotelling's T_0^2 . *Biometrika*, **61**, 381-383.
- RAO C.R., (1973) *Linear Statistical Inference and Its Applications*, 2nd ed. Wiley, New York.
- SCHOTT J.R. (1984), Optimal bounds for the distributions of some test criteria for tests of dimensionality. *Biometrika*, **71**, 561-567.
- SEBER G.A.F. (1984), *Multivariate Observations*. Wiley, New York.
- SEO T., KANDA T., FUJIKOSHI Y. (1995), The effects of non normality on tests for dimensionality in canonical correlation and MANOVA models. *Journal of Multivariate Analysis*, **52**, 325-337.
- TOMASSONE R., DANZART M., DAUDIN J.J., MASSON J.P. (1988), *Discrimination et Classement*, Masson.