REVUE DE STATISTIQUE APPLIQUÉE

F. PIRAUX
R. PALM

Étude empirique des estimateurs des taux d'erreur en analyse discriminante

Revue de statistique appliquée, tome 49, nº 4 (2001), p. 71-85
<http://www.numdam.org/item?id=RSA_2001__49_4_71_0>

© Société française de statistique, 2001, 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/

Rev. Statistique Appliquée, 2001, **XLIX** (4), 71-85

ÉTUDE EMPIRIQUE DES ESTIMATEURS DES TAUX D'ERREUR EN ANALYSE DISCRIMINANTE

F. Piraux[1], R. Palm[2]

[1] *Institut technique des Céréales et des Fourrages, 91720 Boigneville (France)*
[2] *Faculté universitaire des Sciences agronomiques de Gembloux, Avenue de la Faculté d'Agronomie, 8, 5030 Gembloux (Belgique)*

RÉSUMÉ

À partir de simulations, on compare les performances de dix estimateurs paramétriques et six estimateurs non paramétriques dans le cas de l'estimation de trois taux d'erreur en analyse discriminante (taux d'erreur réel, attendu et optimal). Cette comparaison est limitée au cas de l'analyse discriminante linéaire appliquée à deux populations normales de même matrice de variances et covariances. Les simulations ont montré qu'il existe des différences importantes entre les estimateurs et l'utilisateur doit être mis en garde vis-à-vis de certains estimateurs proposés dans les logiciels statistiques.

Mots-clés : *taux d'erreur, classement erroné, fonction discriminante linéaire, bootstrap, jackknife, validation croisée, Monte Carlo.*

ABSTRACT

Monte Carlo experiments are performed to compare ten parametric and six non parametric estimators of the three error rates (actual, expected and optimal error rate) in discriminant analysis. This comparison is restricted to the context of two p-dimensional normal group conditional distributions with the same covariance matrix. The study shows that certain estimators are markedly superior and users should be warned against some estimators available in statistical softwares.

Keywords : *error rates, misclassification, linear discriminant function, bootstrap, jackknife, crossvalidation, Monte Carlo.*

1. Introduction

L'objectif de l'analyse discriminante décisionnelle est de définir une règle permettant de classer un individu dans un groupe particulier, parmi g groupes possibles. Cette affectation se fait sur la base de p variables observées sur cet individu et la règle de classement est établie en fonction des observations de ces mêmes p variables, réalisées sur des échantillons provenant des g groupes ou populations.

Un des problèmes auquel se trouve confronté l'utilisateur de l'analyse discriminante est l'évaluation de la qualité de la règle de décision établie. Cette évaluation

repose souvent sur l'estimation d'un taux d'erreur, c'est-à-dire d'une probabilité de classement erroné.

Plusieurs taux d'erreur, correspondant à des concepts différents, peuvent être définis. Le taux d'erreur optimal, eo, correspond au taux d'erreur quand une règle d'affectation, basée sur les paramètres réels des populations, est appliquée à ces populations. Le taux d'erreur réel ou conditionnel (à un échantillon particulier), ec, est la proportion d'individus mal classés obtenue lorsqu'une règle de classement, basée sur un échantillon particulier prélevé dans chacune des g populations, est appliquée à d'autres individus provenant du même mélange de populations. Enfin, le taux d'erreur réel attendu, ea, est l'espérance mathématique du taux réel pour tous les échantillons d'une taille donnée qui auraient pu être prélevés dans les populations dans le but d'établir la règle de classement. Une discussion de la pertinence de ces taux en fonction des objectifs est donnée dans McLachlan [1992], notamment.

Les expressions analytiques pour le calcul de ces taux d'erreur n'existent que pour des situations particulières et notamment pour le cas de deux populations normales de même matrice de variances et covariances, la règle de classement étant basée sur la fonction linéaire discriminante. Les différents taux d'erreur peuvent, dans ce cas, être exprimés, par des relations exactes ou approchées, en fonction des caractéristiques des deux échantillons et des paramètres des populations.

En pratique, ces paramètres sont le plus souvent inconnus et les taux d'erreur doivent être estimés. De nombreux estimateurs ont été proposés dans ce contexte.

De manière plus générale, pour des nombres et des distributions de populations quelconques et indépendamment de la règle de classement retenue, des estimateurs non paramétriques peuvent être utilisés. Une synthèse des estimateurs disponibles est donnée dans McLachlan [1992], notamment.

L'objectif de notre étude est de comparer différents estimateurs, en nous limitant au cas de l'analyse discriminante linéaire pour deux populations normales, de même matrice de variances et covariances. Cette restriction se justifie pour des raisons pratiques. Pour ce cas on peut, en effet, calculer les trois taux d'erreur théoriques par les formules auxquelles nous avons fait allusion ci-dessus. D'autre part, c'est pour cette situation que les estimateurs sont les plus nombreux.

Pour cette situation, plusieurs auteurs ont comparé les estimateurs à partir de simulations [Ganeshanandam et Krzanowski, 1990; Lachenbruch et Mickey, 1968; McLachlan, 1974a; 1974b; Page, 1985; Snapinn et Knoke, 1984; Sorum, 1972b]. Le nombre d'estimateurs comparés dans ces études, en particulier dans la classe des estimateurs non paramétriques, est plus réduit que dans la présente comparaison. La confrontation de plusieurs estimateurs non paramétriques est particulièrement intéressante, puisque seuls ces estimateurs sont disponibles pour des situations plus générales que celles retenues pour l'étude et, en attendant les résultats d'études plus complètes, on peut supposer, *a priori*, que ces estimateurs auront, dans d'autres situations, des comportements analogues à ceux mis en évidence dans notre comparaison.

Nous présenterons d'abord les estimateurs des taux d'erreur qui seront comparés (paragraphe 2); ensuite nous décrirons le plan de simulation qui a été utilisé (paragraphe 3). Nous examinerons alors les résultats (paragraphe 4) avant de tirer quelques conclusions (paragraphe 5).

2. Taux d'erreur théoriques et estimateurs comparés

2.1. Situation considérée et notations

Nous considérons le cas de deux populations normales à p dimensions, de moyennes égales à $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$ et de matrices de variances et covariances identiques et égales à $\boldsymbol{\Sigma}$. La distance de Mahalonobis séparant ces deux populations est égale à :

$$\Delta = \left[(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)' \, \boldsymbol{\Sigma}^{-1} \, (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \right]^{1/2} .$$

Soit un échantillon d'effectif n prélevé dans chacune de ces populations. Soient $\bar{\boldsymbol{x}}_1$ et $\bar{\boldsymbol{x}}_2$, les deux vecteurs des moyennes de ces deux échantillons et $\boldsymbol{S}$ la matrice de variances et covariances commune estimée. La règle de classement est la suivante : un individu, caractérisé par un vecteur d'observations $\boldsymbol{x}$, est classé dans la population 1 si $y(\boldsymbol{x}) \geqslant 0$ et dans la population 2 si $y(\boldsymbol{x}) < 0$, $y(\boldsymbol{x})$ étant la fonction linéaire discriminante calculée au point $\boldsymbol{x}$ et définie par :

$$y(\boldsymbol{x}) = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' \, \boldsymbol{S}^{-1} \left[\boldsymbol{x} - \frac{1}{2} (\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \right] .$$

Pour la situation envisagée et en considérant des probabilités *a priori* d'appartenance aux deux populations égales, la règle de discrimination retenue, appelée règle de discrimination linéaire conduit au taux d'erreur attendu minimum.

2.2. Taux d'erreur théoriques

Le taux d'erreur optimal est égal à [Huberty, 1994; McLachlan, 1992] :

$$eo = \Phi(-\Delta/2) ,$$

Φ représentant la fonction de répartition de la variable normale réduite.

Le taux d'erreur réel spécifique au groupe i est donné par [McLachlan, 1975] :

$$ec_i = \Phi \left[(-1)^i \frac{[\boldsymbol{\mu}_i - \frac{1}{2}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2)] \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)}{[(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2) \boldsymbol{S}^{-1} \boldsymbol{\Sigma} \boldsymbol{S}^{-1} (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}} \right] \quad (i = 1, 2) ,$$

et le taux d'erreur réel général théorique est égal à :

$$ec = (ec_1 + ec_2)/2 .$$

Quant au taux d'erreur attendu, ea, son expression analytique est très compliquée [John, 1961]. Une relation approchée est donnée par McLachlan [1992] et

Sorum [1972a] :

$$ea = \Phi(-\Delta/2) + \frac{1}{8}\phi(\Delta/2)[4(p-1)\Delta^{-1} + p\Delta]n^{-1}\,,$$

Φ et ϕ représentant respectivement la fonction de répartition et la fonction de densité de probabilité de la variable normale réduite.

2.3. Estimateurs paramétriques

Plusieurs estimateurs paramétriques ont comme point de départ la relation donnant le taux d'erreur optimal théorique. Ils se différencient par l'expression utilisée à la place de Δ. En désignant par D la distance de Mahalanobis séparant les deux échantillons :

$$D = [(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)'\,\boldsymbol{S}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)]^{1/2}\,,$$

les estimateurs suivants ont été considérés :

$$eD = \Phi(-D/2)\,,$$

$$eDS = \Phi\left[-\frac{D}{2}\left(\frac{2\,n-p-3}{2\,n-2}\right)^{1/2}\right]\,,$$

$$e3 = \Phi\left[-\frac{D}{2}(1+\frac{1}{n})^{-1/2}\right]\,,$$

$$e5 = \Phi\left[-\frac{D}{2}\left(1+\frac{1}{2\,n}\right)^{-1/2}\right]\,,$$

$$eB = \Phi\left[-\frac{D}{2}\left(\frac{2\,n(2\,n-p-4}{(2\,n-1)^2}\right)^{1/2}\right]\,.$$

Ces estimateurs ont été étudiés par Sorum [1972b], qui donne des références sur leur origine.

Deux autres estimateurs testés correspondent à la formule du taux d'erreur attendu théorique, ea, après remplacement de Δ, soit par D, soit par :

$$DS = D[(2\,n-p-3)/(2\,n-2)]^{1/2}\,.$$

Ils sont désignés par eO et eOS.

L'estimateur eM a été proposé par McLachlan [1975] pour l'estimation du taux réel, ec, dans le but de réduire le biais. Cette expression étant très complexe, nous ne

la reproduisons pas ici; elle peut être trouvée également dans McLachlan [1992] et dans Huberty [1994].

L'estimateur eL a été proposé par Lachenbruch [1968]. Il s'écrit :

$$eL = \Phi\left[-M^{-1/2}[\frac{1}{2}D - (p\,N\,D^{-1})/n(N-p-1)]\right],$$

avec : $N = 2(n-1)$ et $M = N(N-1)/(N-p)(N-p-3)$.

Enfin, l'estimateur eU, défini par Lachenbruch et Mickey [1968], combine la méthode de validation croisée et les caractéristiques d'une méthode paramétrique. Chaque observation est éliminée tour à tour des données et la fonction linéaire discriminante, $y(\boldsymbol{x})$, est déterminée sur les $2n-1$ observations. La valeur de cette fonction est alors calculée pour l'observation éliminée. La moyenne et l'écart-type sont ensuite déterminés pour chaque échantillon. Soit $\bar{y}_1$ et s_{y_1} la moyenne et l'écart-type des valeurs obtenues pour le premier échantillon et $\bar{y}_2$ et s_{y_2} la moyenne et l'écart-type des valeurs obtenues pour le deuxième échantillon. On a :

$$eU = \frac{1}{2}\left[\Phi(-\bar{y}_1/s_{y_1}) + \Phi(\bar{y}_2/s_{y_2})\right].$$

Les estimateurs eL et eU, sont, comme eM, des estimateurs du taux réel, ec.

2.4. Estimateurs non paramétriques

Le taux de resubstitution, appelé aussi taux d'erreur apparent, eA, est la proportion d'individus mal classés quand on utilise la règle de classement établie sur les $2n$ individus pour reclasser ces mêmes individus.

Le taux d'erreur par validation croisée, eCV, est la proportion d'individus mal classés lorsqu'on effectue $2n$ fois l'analyse discriminante sur les $2n-1$ données, en mettant de côté tout à tour chacune des observations et en reclassant cette observation sur la base de la règle établie sur les $2n-1$ autres données.

L'estimateur du *Jackknife*, eJc, est obtenu, comme l'estimateur eCV, par la réalisation de $2n$ analyses discriminantes sur $2n-1$ observations. Pour chaque échantillon de $2n-1$ observations, l'observation i étant éliminée, les taux d'erreur apparents par groupe, $eA_{1(i)}$ et $eA_{2(i)}$, sont calculés. Soient $\bar{eA}_1$ et $\bar{eA}_2$ les moyennes de ces taux :

$$\bar{eA}_1 = \frac{1}{n}\sum_{i=1}^{n} eA_{1(i)} \quad \text{et} \quad \bar{eA}_2 = \frac{1}{n}\sum_{i=1}^{n} eA_{2(i)}$$

et eA_1 et eA_2 les taux d'erreur apparents des groupes pour l'échantillon complet. L'estimateur s'écrit alors [McLachlan, 1992] :

$$eJc = \frac{1}{2}\left[\left(eA_1 + (n-1)(eA_1 - \bar{eA}_1)\right) + \left(eA_2 + (n-1)(eA_2 - \bar{eA}_2)\right)\right].$$

L'estimation par la méthode du *bootstrap*, $eBoot$, a été obtenue par 100 rééchantillonnages, un échantillon de taille n étant prélevé de manière aléatoire et simple, avec remise, dans chacun des échantillons initiaux de n observations. Pour chaque couple d'échantillons bootstrapés, la règle de classement est établie et utilisée pour reclasser les individus de l'échantillon bootstrapé et définir les taux d'erreur apparents par groupe, eA^*_{1j} et eA^*_{2j} $(j = 1, \ldots, 100)$. Cette règle de classement est également utilisée pour reclasser les observations des échantillons initiaux et définir les taux d'erreur apparents par groupe, eA^{**}_{1j} et eA^{**}_{2j} $(j = 1, \ldots, 100)$. En désignant par b_1 et b_2 les quantités :

$$b_1 = \frac{1}{100}\sum_{j=1}^{100}(eA^*_{1j} - eA^{**}_{1j}) \quad \text{et} \quad b_2 = \frac{1}{100}\sum_{j=1}^{100}(eA^*_{2j} - eA^{**}_{2j}),$$

le taux d'erreur estimé par *bootstrap* s'écrit :

$$eBoot = [(eA_1 - b_1) + (eA_2 - b_2)]/2,$$

eA_1 et eA_2 étant les taux d'erreur apparents sur les n individus initiaux de chaque groupe, lorsque ceux-ci sont reclassés à partir de la règle établie sur les $2n$ observations initiales [McLachlan, 1992].

L'estimateur ePP se base sur les probabilités d'appartenance *a posteriori* au groupe lorsqu'on reclasse les individus de l'échantillon, comme dans le cas du taux d'erreur apparent. Si $\widehat{\tau}_{1i}$ et $\widehat{\tau}_{2i}$ représentent les probabilités *a posteriori* d'appartenance de l'individu i au groupe 1 et au groupe 2, on a, pour des probabilités *a priori* égales :

$$\widehat{\tau}_{1i} = \widehat{f}_1(\boldsymbol{x}_i)/\left[\widehat{f}_1(\boldsymbol{x}_i) + \widehat{f}_2(\boldsymbol{x}_i)\right] \quad \text{et} \quad \widehat{\tau}_{2i} = \widehat{f}_2(\boldsymbol{x}_i)/\left[\widehat{f}_1(\boldsymbol{x}_i) + \widehat{f}_2(\boldsymbol{x}_i)\right]$$

$\widehat{f}_1(\boldsymbol{x}_i)$ et $\widehat{f}_2(\boldsymbol{x}_i)$ étant les densités de probabilité estimées relatives aux deux populations. L'estimateur ePP s'écrit alors :

$$ePP = 1 - \frac{1}{2n}\sum_{i=1}^{2n}\max(\widehat{\tau}_{1i}, \widehat{\tau}_{2i}),$$

max $(\widehat{\tau}_{1i}, \widehat{\tau}_{2i})$ représentant le maximum de $\widehat{\tau}_{1i}$ et $\widehat{\tau}_{2i}$.

L'estimateur $ePPCV$ se définit de la même manière que l'estimateur ePP, mais les probabilités *a posteriori* sont obtenues à partir de la règle de classement établie sur $2n - 1$ individus, l'individu i étant exclu.

3. Simulations

3.1. Facteurs contrôlés

Pour un taux d'erreur donné, l'efficacité des différents estimateurs étudiés varie en fonction du nombre p de variables, de la taille des deux échantillons et de la différence Δ qui existe entre les deux populations.

Pour p, les trois valeurs suivantes ont été retenues : 4, 8 et 16. Pour la taille des échantillons, nous avons considéré uniquement des effectifs identiques pour les deux échantillons ($n_1 = n_2 = n$) et, de plus, la valeur de n a été définie en fonction de p : nous avons considéré, d'une part, un rapport p/n égal à 0,4 et, d'autre part, un rapport p/n égal à 0,2. Enfin, les trois distances Δ retenues ont été définies de manière à obtenir un taux optimal théorique respectivement de 10 % ($\Delta \simeq 2,56$), de 20 % ($\Delta \simeq 1,68$) et de 30 % ($\Delta \simeq 1,05$).

3.2. Génération des données

Si $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ sont deux variables aléatoires normales à p dimensions, de moyennes $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$ et de matrice de variances et covariances égale à $\boldsymbol{\Sigma}$, les taux d'erreur théoriques seront identiques à ceux des variables aléatoires $\boldsymbol{y}_1$ et $\boldsymbol{y}_2$, obtenues par une même transformation linéaire de $\boldsymbol{x}_1$ et $\boldsymbol{x}_2$ [Dunn, 1971] :

$$\boldsymbol{y}_1 = \boldsymbol{A}\boldsymbol{x}_1 + \boldsymbol{c} \quad \text{et} \quad \boldsymbol{y}_2 = \boldsymbol{A}\boldsymbol{x}_2 + \boldsymbol{c}\,.$$

Il en résulte que, pour p et n fixés, les taux d'erreur sont uniquement fonction de Δ. Pour tenir compte de Δ, on peut donc, sans perte de généralité, simuler une première population normale dont les paramètres sont $\boldsymbol{\mu}_1 = \boldsymbol{0}$ et $\boldsymbol{\Sigma} = \boldsymbol{I}$ et une deuxième population normale dont les paramètres sont $\boldsymbol{\mu}_2 = (\Delta\ 0 \ldots 0)'$ et $\boldsymbol{\Sigma} = \boldsymbol{I}$ [Lachenbruch et Mickey, 1968]. La génération des données peut, par conséquent, se réduire à la génération de variables normales indépendantes.

3.3. Critères de comparaison

Au total, 18 situations résultant de la combinaison de trois valeurs de p, deux tailles d'échantillons et trois distances ont donc été considérées. Pour chaque situation, on a généré 100 couples d'échantillons. Pour chaque couple, on a déterminé :

- la valeur des 16 estimateurs présentés aux paragraphes 2.3 et 2.4;
- la valeur théorique des trois taux d'erreur par les relations du paragraphe 2.2. Pour le taux d'erreur optimal et pour le taux d'erreur attendu, les 100 valeurs relatives à une même situation sont identiques, puisque ces taux ne dépendent pas des paramètres des échantillons.

Ensuite, pour chaque situation, on a déterminé l'erreur absolue moyenne entre les estimations et les taux théoriques ainsi que le biais :

$$EAM = \frac{1}{100} \sum_{i=1}^{100} | \text{ taux estimé - taux théorique } |$$

et

$$\text{biais} = \frac{1}{100} \sum_{i=1}^{100} (\text{taux estimé - taux théorique}) .$$

Les valeurs de EAM et du biais ont été déterminées pour chacun des estimateurs présentés au paragraphe 2 et pour chacun des trois taux théoriques. Ainsi donc, chacun des 16 estimateurs est considéré, dans un premier temps, comme un estimateur de eo, ensuite comme un estimateur de ec et, enfin, comme un estimateur de ea.

4. Résultats et discussion

4.1. Présentation des résultats

Les erreurs absolues moyennes varient de manière relativement importante pour les différentes situations envisagées. Elles sont, par exemple, plus importantes lorsque l'effectif est faible; elles sont plus importantes aussi lorsque le taux d'erreur optimal est grand, c'est-à-dire lorsque Δ est faible.

L'objectif de l'étude n'étant pas de mettre en évidence l'incidence de ces facteurs sur l'ordre de grandeur des erreurs absolues moyennes, mais de comparer des estimateurs dans des situations variées, nous avons remplacé les erreurs absolues moyennes par des rangs. Pour un taux d'erreur donné et pour une combinaison des facteurs étudiés, le rang de chacun des estimateurs a donc été déterminé, l'estimateur présentant l'erreur absolue moyenne la plus faible obtenant le rang 1. Les 18 rangs obtenus par chacun des estimateurs ont ensuite été représentés sous la forme d'un graphique en boîtes multiples établi par le logiciel Minitab [Tukey, 1977; X, 1996]. Ce graphique donne le classement des 16 estimateurs en fonction de la médiane des rangs.

Ensuite, pour quantifier l'importance des différences entre estimateurs, nous avons exprimé les erreurs absolues moyennes en proportion de l'erreur absolue moyenne obtenue, pour la même situation, par l'estimateur classé globalement en première position. Les résultats ont été présentés sous la forme d'un graphique en boîtes multiples.

Enfin, un graphique en boîtes multiples a encore été établi pour représenter la distribution des biais, dans le but d'expliquer, dans la mesure du possible, les erreurs absolues moyennes importantes produites par certains estimateurs.

Nous avons également vérifié si d'éventuelles interactions entre les estimateurs et les facteurs contrôlés (n, p et Δ), qui seraient masquées dans les représentations graphiques décrites ci-dessus devaient être prises en considération dans l'interprétation des résultats.

4.2. Estimation du taux d'erreur réel

La distribution des rangs (figure 1) montre que les meilleurs résultats sont obtenus pour l'estimateur eOS qui est pourtant un estimateur du taux d'erreur attendu. Celui-ci est en effet classé en première position dans 10 cas sur 18, en deuxième position dans 3 cas sur 18 et son plus mauvais classement correspond au rang 7.

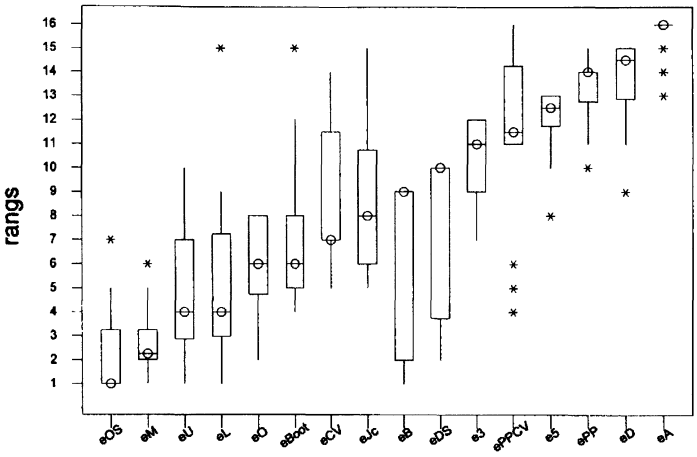

FIGURE 1

Estimation du taux d'erreur réel : distribution des rangs (les méthodes sont classées par ordre croissant du rang médian).

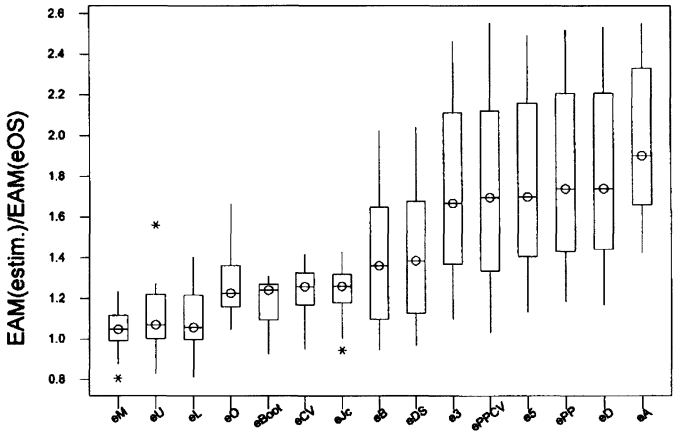

FIGURE 2

Estimation du taux d'erreur réel : distribution des erreurs absolues moyennes, exprimées en proportion de l'erreur absolue moyenne de la méthode eOS.

Les estimateurs eU, eL et eM se classent également bien, l'estimateur eU, plus compliqué à mettre en oeuvre, ne présentant pas d'avantages particuliers. Le bon comportement de ces trois estimateurs est assez logique, dans la mesure où ils ont précisément été proposés pour estimer le taux d'erreur réel.

A l'opposé, on note les très mauvaises performances des estimateurs eA et eD. La validation croisée est très nettement supérieures à eA et les corrections apportées à l'estimation de Δ améliorent effectivement les performances des estimateurs basés sur ces corrections : eB, eDS, $e3$ et $e5$ sont légèrement meilleurs que eD.

Les estimateurs non paramétriques, $eBoot$, eCV et eJc, présentent globalement un comportement assez proche, avec toutefois un léger avantage pour $eBoot$, qui se montre supérieur à eJc dans 13 cas sur 18 et supérieur à eCV dans 15 cas sur 18. Par ailleurs, pour la validation croisée, la prise en compte des probabilités *a posteriori* n'améliore pas la qualité des estimateurs, $ePPCV$ étant moins bien classé que eCV dans 14 cas sur 18.

Le facteur affectant, dans l'ensemble, le plus le rang obtenu par les estimateurs est le nombre de variables p. Les estimateurs eCV, eJc, $eBoot$, eM, eL et eU voient leur classement s'améliorer lorsque p augmente alors que, au contraire, les estimateurs eD, eDS, $e3$, $e5$, eB, eO, ePP et $ePPCV$ obtiennent alors un moins bon classement.

La figure 2 met en évidence une perte d'efficacité par rapport à l'estimateur eOS, de l'ordre de 5 à 10 % pour eM, eL et eU, de 20 à 30 % pour $eBoot$, eCV et eJc, de 30 à 40 % pour eB et eDS et une perte supérieure à 90 % pour les autres méthodes. De manière générale, la perte d'efficacité est plus faible lorsque la taille de l'échantillon augmente, sauf pour $eBoot$, eCV et eJc, pour lesquels la taille n'a pas d'effet. La perte d'efficacité augmente par contre avec p sauf pour les trois estimateurs non paramétriques cités ci-dessus et pour les estimateurs eM, eL et eU, pour lesquels elle diminue. Quant à l'effet de Δ, il est moins marqué. Pour eD, eDS, $e3$, $e5$, ePP et $ePPCV$, la perte d'efficacité est plus faible lorsque Δ augmente; pour $eBoot$, elle est plus grande quand Δ augmente et pour les autres estimateurs, la perte d'efficacité n'est pas liée à Δ.

L'examen des biais (figure 3) permet d'expliquer, dans une large mesure, le mauvais comportement de certains estimateurs. On constate, en effet, que les estimateurs occupant les huit dernières positions présentent un biais négatif très important : quelle que soit la situation envisagée, ils conduisent à une sous-estimation du taux d'erreur, qui peut atteindre près de 15 % dans les situations les plus défavorables, caractérisées par un effectif faible, un nombre de variables élevé et une distance Δ faible.

On note aussi le biais pratiquement nul des méthodes non paramétriques $eBoot$, eCV et eJc et des estimateurs eM et eOS, ce dernier présentant un biais négatif de 5 % dans une seule situation.

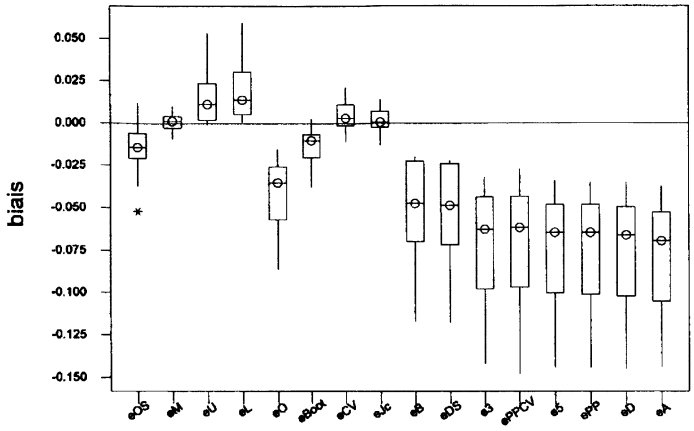

FIGURE 3
Estimation du taux d'erreur réel : distribution des biais.

4.3. Estimation du taux d'erreur attendu

L'estimation du taux attendu conduit, dans l'ensemble, aux mêmes constatations que celles émises pour le taux d'erreur réel : le classement des méthodes à partir du rang médian est identique à celui de la figure 1. Les rapports des erreurs absolues moyennes des méthodes à l'erreur moyenne de la méthode eOS donnent lieu à une représentation graphique comparable à la figure 2 et les biais se présentent comme dans la figure 3. Pour cette raison, nous ne reprenons pas ici ces graphiques.

Rappelons que eO et eOS ont été définis pour estimer le taux d'erreur attendu. Leur bon classement est donc relativement logique.

4.4. Estimation du taux d'erreur optimal

La figure 4 montre que le classement des estimateurs est assez différent de celui obtenu pour le taux réel et le taux espéré. L'estimateur eB obtient globalement la première position : il a donné le meilleur résultat dans 9 cas sur 18 et, dans le cas le plus défavorable, il se situe en cinquième position.

Les estimateurs basés sur les distances dérivées de D (eD, eDS, $e3$, $e5$ et eB) et destinés à estimer le taux d'erreur optimal se classent, dans l'ensemble, mieux que les estimateurs initialement prévus pour l'estimation du taux réel (eU, eM et eL). Les estimateurs eO et eOS se classent relativement bien également .

Enfin, les méthodes non paramétriques obtiennent également un mauvais classement.

La figure 5 montre que eB et eDS ont des efficacités très proches et que eO n'est que légèrement inférieur aux deux autres estimateurs.

L'erreur absolue moyenne relativement élevée liée à eCV, eJc, $eBoot$, eM, eL et eU résulte de la surestimation du taux d'erreur optimal, qui s'observe pour les 18 situations (figure 6). Par contre, les estimateurs eA, ePP, $ePPCV$, eD, $e3$ et $e5$ donnent lieu à une sous-estimation dans les 18 cas.

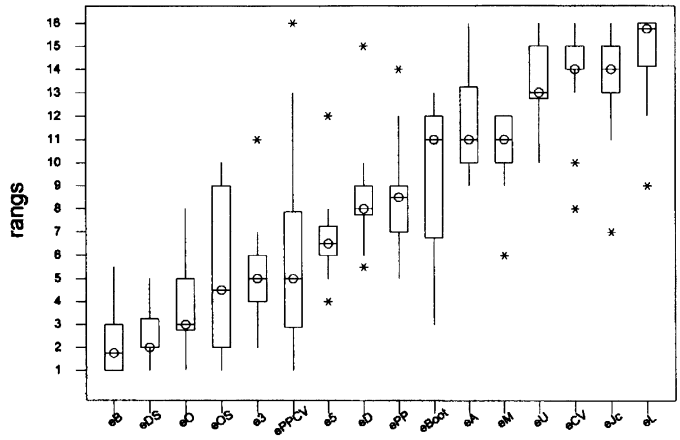

FIGURE 4

Estimation du taux d'erreur optimal : distribution des rangs (les méthodes sont classées par ordre croissant du rang médian).

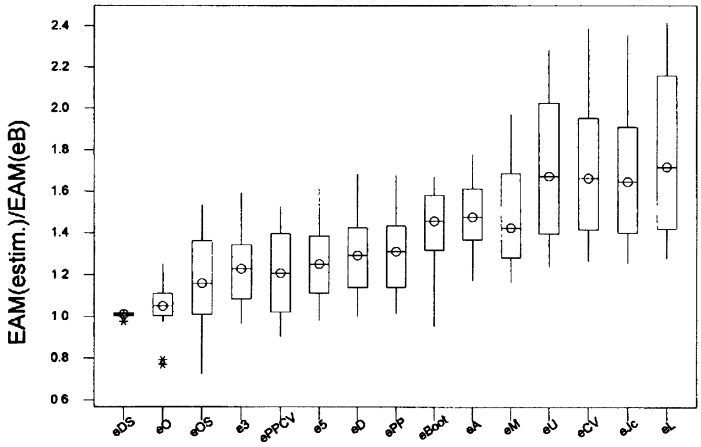

FIGURE 5

Estimation du taux d'erreur optimal : distribution des erreurs absolues moyennes, exprimées en proportion de l'erreur absolue moyenne de la méthode eB.

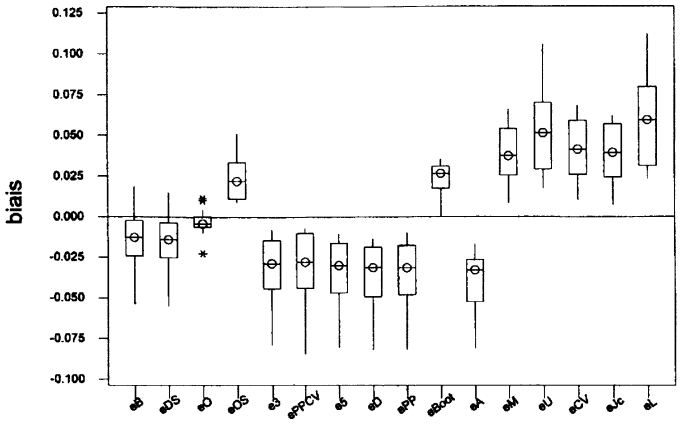

FIGURE 6
Estimation du taux d'erreur optimal : distribution des biais.

5. Conclusions

Les simulations réalisées montrent que des différences importantes existent entre les estimateurs étudiés, les estimateurs les plus mauvais conduisant, en moyenne, à des erreurs moyennes absolues de 1,5 à 2 fois supérieures à celles obtenues pour les estimateurs les meilleurs. Aucun estimateur ne s'est montré supérieur aux autres dans toutes les situations étudiées, mais il ne semble pas possible de nuancer les recommandations pratiques en fonction de la situation considérée, c'est-à-dire en fonction du nombre d'observations dans les échantillons, du nombre de variables et de la distance entre les deux populations.

Le classement des estimateurs est tout a fait similaire pour l'estimation du taux d'erreur réel et du taux d'erreur attendu. Par contre, le classement est assez différent lorsqu'on estime le taux d'erreur optimal. Pour le taux d'erreur réel et le taux d'erreur attendu, l'estimateur eOS s'est avéré le meilleur en moyenne. Les estimateurs eL, eM et eU lui sont légèrement inférieurs. Pour le taux d'erreur optimal, les meilleurs estimateurs sont eB et eDS.

Les conclusions rejoignent globalement, pour le taux d'erreur réel, celles de Page [1985] en ce qui concerne la supériorité de eL et eM et eOS par rapport à eD, eDS, $e3$, $e5$ et eO. Elles rejoignent également celles de Lachenbruch et Mickey [1968], qui ont montré la supériorité de eOS, sauf si Δ est petit et si les effectifs sont faibles, ainsi que les conclusions de Ganeshanandam et Krzanowski [1990] qui, pour 11 estimateurs testés, ont obtenu l'ordre décroissant suivant, pour les meilleurs estimateurs : eM, eL, eU, eOS et eCV. La supériorité de certains estimateurs paramétriques par rapport aux estimateurs non paramétriques a également été mise en évidence par plusieurs auteurs [Snapinn et Knoke, 1984; Sorum, 1972b].

Pour le taux d'erreur optimal, Sorum [1972b] a également trouvé que les estimateurs eDs, eB et eO sont les meilleurs parmi les estimateurs testés dans son étude.

Les simulations montrent aussi que l'utilisateur doit être mis en garde vis-à-vis des estimateurs proposés dans les logiciels tels que Minitab et SAS. Le premier propose eA et eCV et le second propose, en plus, ePP et $ePPCV$ [SAS, 1989; X, 1996]. Parmi ces quatre estimateurs, trois se sont montrés particulièrement mauvais pour l'estimation du taux d'erreur réel (eA, ePP et $ePPCV$). Par contre, l'estimateur eCV occupe un rang intermédiaire, puisqu'il se situe en septième position, l'errreur absolue moyenne étant, en moyenne, 1,25 fois plus grande que celle de eOS. Pour l'estimation du taux d'erreur optimal, par contre, les estimateurs eA et eCV sont particulièrement mauvais et, parmi les estimateurs disponibles dans SAS, $ePPCV$ est le meilleur. Dans notre étude, il se classe en sixième position.

Enfin, bien qu'on ait considéré 18 situations différentes, l'étude est limitée au cas de deux populations normales de même matrice de variances et covariances, d'échantillons de même effectif, le classement étant réalisé par la fonction linéaire discriminante (paragraphe 2.1). Il est difficile de prévoir le comportement de ces estimateurs lorsqu'on s'écarte de cette situation. Pour le taux d'erreur réel et attendu, le comportement satisfaisant de eCV, eJc et $eBoot$ dans les conditions de notre étude nous suggère de recommander ces estimateurs pour des situations autres que celles envisagées ci-dessus. On peut, en effet, penser *a priori* que la qualité de ces estimateurs est moins fonction des caractéristiques des populations que ne l'est la qualité des estimateurs paramétriques. Le choix de l'un ou l'autre parmi ces trois estimateurs nous paraît relativement accessoire, malgré le léger avantage de $eBoot$ sur eCV et eJc.

Bibliographie

DUNN O.J. [1971]. Some expected values for probabilities of correct classification in discriminant analysis. *Technometrics* 13, 345-353.

GANESHANANDAM S., KRZANOWSKI W.J. [1990]. Error-rate estimation in two-group discriminant analysis using the linear discriminant function. *J. Stat. Comput. Simul.* 36, 157-175.

HUBERTY C.J. [1994]. *Applied discriminant analysis.* New York, Wiley, 466 p.

JOHN S. [1961]. Errors in discrimination. *Ann. Math. Stat.* 32, 1125-1144.

LACHENBRUCH P.A. [1968]. On expected probabilities of misclassification in discriminant analysis, necessary sample size, and a relation with the multiple correlation coefficient. *Biometrics* 24, 823-834.

LACHENBRUCH P.A., MICKEY M.R. [1968]. Estimation of error rates in discriminant analysis. *Technometrics* 10,1-11.

MCLACHLAN G.J. [1974a]. Estimation of the errors of misclassification on the criterion of asymptotic mean square error. *Technometrics* 16, 255-260.

MCLACHLAN G.J. [1974b]. An asymptotic unbiased technique for estimating the error rates in discriminant analysis. *Biometrics* 30, 239-249.

MCLACHLAN G.J. [1975]. Confidence intervals for the conditional probability of misallocation in discriminant analysis. *Biometrics* 31, 161-167.

MCLACHLAN G.J. [1992]. *Discriminant analysis and statistical pattern recognition.* New York, Wiley, 526 p.

PAGE J.T. [1985]. Error-rate estimation in discriminant analysis. *Technometrics* 27, 189-198.

SAS INSTITUTE INC [1989]. *SAS/STAT User's guide, version 6*, Fourth edition (2 volumes). Cary NC : SAS Institute Inc., 943 + 946 p.

SNAPINN S.M., KNOKE J.D. [1984]. Classification error-rate estimators evaluated by unconditional mean squared error. *Technometrics* 26, 371-378.

SORUM M. [1972a]. Three probabilities of misclassification. *Technometrics* 14, 309-316.

SORUM M. [1972b]. Estimating the expected and optimal probabilities of misclassification. *Technometrics* 14, 935-943.

TUKEY J. [1977]. *Exploratory data analysis.* Reading, Addison-Wesley, 688 p.

X [1996]. *Minitab reference manual : release 11 for Windows.* PA State College, Minitab, 1052 p.