REVUE DE STATISTIQUE APPLIQUÉE

R. GLÈLÈ KAKAÏ
F. PIRAUX
N. H. FONTON
R. PALM

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

Revue de statistique appliquée, tome 51, n° 3 (2003), p. 61-74
<http://www.numdam.org/item?id=RSA_2003__51_3_61_0>

Rev. Statistique Appliquée, 2003, **LI** (3), 61-74

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

R. GLÈLÈ KAKAÏ[1], F. PIRAUX[2], N.H. FONTON[1], R. PALM[3]

[1] *Faculté des Sciences Agronomiques, Université d'Abomey-Calavi, BP 526, Cotonou (Bénin)*

[2] *Institut technique des Céréales et des Fourrages, 91720 Boigneville (France)*

[3] *Faculté universitaire des Sciences agronomiques de Gembloux, Avenue de la Faculté d'Agronomie, 8, 5030 Gembloux*

RÉSUMÉ

A partir de simulations, on compare les performances de vingt estimateurs des taux d'erreur en analyse discriminante, dans le cas de deux populations et pour la règle de discrimination linéaire. Différentes distributions (normale, chi-2 et bêta) ont été considérées. On conclut que l'estimateur eOS (pour le taux réel et le taux attendu) et eB (pour le taux optimal) sont les estimateurs les meilleurs, sauf pour des distributions s'écartant très nettement des populations normales. L'estimateur $e632$ est le meilleur estimateur non paramétrique. Il est préférable aux estimateurs paramétriques si les distributions sont très différentes des distributions normales.

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

ABSTRACT

Monte Carlo experiments are performed to compare twenty estimators of error rates in discriminant analysis in the case of two populations and linear classification rule. Several distributions (normal, chi-square and beta) have been considered. Estimators eOS (for actual and expected error rate) and eB (for optimal error rate) are the best, except for distributions very different from normal distributions. The $e632$ estimator is the best non parametric estimator. This estimator is better than parametric estimators for distributions very different from normal distributions.

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

1. Introduction

Dans une étude antérieure, Piraux et Palm [2001] ont comparé les performances de dix estimateurs paramétriques et de six estimateurs non paramétriques, dans le cas de l'estimation des taux d'erreur en analyse discriminante linéaire. L'étude

était limitée au cas de deux populations normales de même matrice de variances et covariances.

La présente étude a comme objectif de vérifier dans quelle mesure les conclusions obtenues peuvent être étendues au cas de populations non normales de même matrice de variances et covariances. Nous considérons toujours l'analyse discriminante linéaire et des probabilités *a priori* égales.

Certaines populations considérées s'écartent très nettement des populations normales. La prise en compte de situations relativement extrêmes a pour but de permettre une meilleure appréciation de l'influence de la non-normalité sur le classement des estimateurs des taux d'erreurs; cela ne signifie pas que l'utilisation de l'analyse discriminante linéaire doit être conseillée dans de telles situations.

Nous présentons d'abord la méthode utilisée pour déterminer les différents taux d'erreur théoriques (paragraphe 2); ensuite nous décrivons le plan de simulation adopté et les estimateurs comparés (paragraphe 3). Nous examinons alors les résultats obtenus (paragraphe 4), avant de tirer quelques conclusions (paragraphe 5).

Dans la mesure où la présente étude est une extension d'un travail antérieur, différents aspects ne seront plus détaillés ici mais nous renverrons directement le lecteur au paragraphe concerné de l'article de Piraux et Palm [2001], désigné ci-après par le symbole PP.

2. Taux d'erreur théoriques

2.1. Taux d'erreur optimal théorique

Nous définissons le taux d'erreur optimal, eo, comme étant le taux d'erreur associé à la règle de discrimination linéaire, lorsque celle-ci est déterminée en fonction des paramètres réels des populations et est appliquée à ces mêmes populations. Cette règle s'écrit :

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

un individu caractérisé par un vecteur d'observations $\boldsymbol{x}$ étant classé dans la population 1 si $y(\boldsymbol{x}) \geqslant 0$ et dans la population 2 si $y(\boldsymbol{x}) < 0$. Dans cette relation, $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$ sont les vecteurs des moyennes des deux populations et $\boldsymbol{\Sigma}$ est la matrice des variances et covariances, identique pour les deux populations.

Pour deux populations normales de même matrice de variances et covariances, la règle de discrimination linéaire est la règle optimale et le taux eo est fonction de la distance Δ de Mahalanobis qui sépare les deux populations (PP, paragraphe 2) :

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

Si les populations sont non normales, par contre, la règle de classement basée sur la fonction linéaire n'est plus nécessairement optimale et le taux d'erreur théorique lié à l'utilisation de cette fonction linéaire ne peut plus, d'une manière générale, être

directement exprimé en fonction des deux vecteurs de moyennes $\boldsymbol{\mu}_1$ et $\boldsymbol{\mu}_2$ et de la matrice de variances et covariances commune $\boldsymbol{\Sigma}$ des populations. Il a été obtenu par simulations. Deux échantillons de 10.000 individus, un échantillon provenant de chacune des deux populations, ont été générés et la règle de classement a été déterminée en remplaçant, dans l'expression donnant $y(\boldsymbol{x})$, les paramètres théoriques $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ et $\boldsymbol{\Sigma}$ par leur estimation. Cette règle a ensuite été appliquée à deux autres échantillons de 10.000 individus tirés dans les mêmes populations. Des répétitions ont montré que la taille retenue pour ces échantillons conduit à une estimation du taux d'erreur optimal théorique dont l'erreur-standard est, pour les situations envisagées, de l'ordre de 0,2 à 0,3 pour cent.

2.2. Taux d'erreur réel théorique

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 deux populations est appliquée à d'autres individus provenant du même mélange des deux populations. Ce taux est celui qui est en général le plus utile dans la pratique.

Pour deux populations normales de même matrice de variances et covariances, le taux d'erreur peut être exprimé par une formule en fonction, d'une part, des paramètres $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$ et $\boldsymbol{\Sigma}$ des deux populations, et d'autre part, des moyennes $\bar{\boldsymbol{x}}_1$, $\bar{\boldsymbol{x}}_2$ des deux échantillons et de leur matrice de variances et covariances commune estimée $\boldsymbol{S}$ (PP, paragraphe 2.2).

Pour des populations non normales, ce taux peut être estimé en appliquant la règle de discrimination linéaire construite sur les deux échantillons donnés à un échantillon-test de grande taille. Comme pour le taux optimal, l'échantillon-test est constitué de 10.000 individus de la première population et 10.000 individus de la deuxième population. La précision de cette estimation est du même ordre de grandeur que celle de l'estimation du taux d'erreur optimal.

2.3. Taux d'erreur attendu théorique

Ce taux, noté ea, est l'espérance mathématique du taux réel pour tous les échantillons d'une taille donnée qui pourraient être prélevés dans les populations dans le but d'établir la règle de classement. Une expression analytique de ce taux existe uniquement pour deux populations normales de même matrice de variances et covariances (PP, paragraphe 2.2).

Ce taux peut être estimé en calculant la moyenne de r taux d'erreurs réels théoriques. La valeur de r a été fixée à 100, ce qui conduit à une estimation dont la précision varie de 0,2 à 0,8 pour cent, pour les degrés de recouvrement des deux populations qui ont été considérés.

3. Simulations

3.1. Génération des données

Le principe de génération des données est identique à celui utilisé antérieurement (PP, paragraphe 3.2). Pour la première population, on génère p variables aléatoires indépendantes, de moyenne nulle et d'écart-type unitaire. Pour la deuxième population, on génère $p-1$ variables aléatoires indépendantes de moyenne nulle et d'écart-type unitaire et une $p^{ième}$ variable aléatoire, indépendante des autres, d'écart-type unitaire mais de moyenne différente de zéro, la moyenne étant déterminée de manière à obtenir, pour cette variable, un taux de recouvrement des deux populations fixé *a priori*.

3.2. Facteurs contrôlés

Les facteurs pris en compte sont le nombre p de variables, la taille n des échantillons, la nature des distributions et le taux de recouvrement des 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.

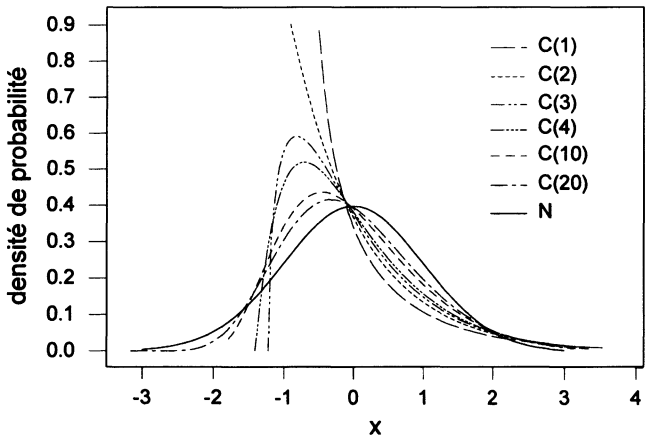

FIGURE 1
Distributions χ^2 et distribution normale

Pour les distributions, 19 situations ont été retenues : la distribution normale (notée N), 8 distributions χ^2 (notées $C(k)$), 6 distributions bêta (notées $B(k_1, k_2)$) et 4 mélanges de distributions. Les distributions χ^2 et bêta ont subi une transformation linéaire, de manière à obtenir un écart-type unitaire et une moyenne fixée. Les degrés

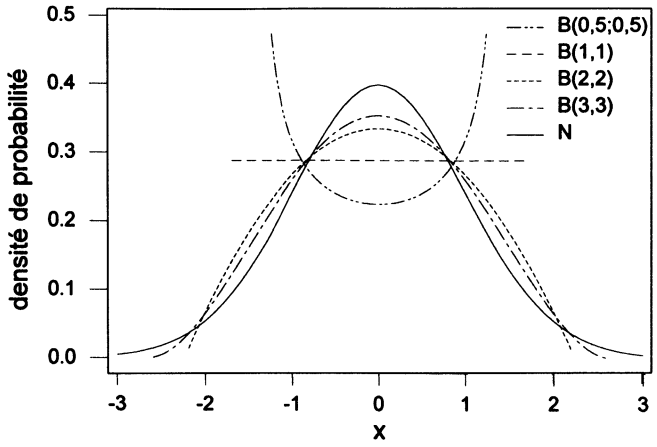

FIGURE 2
Distributions bêta et distribution normale

de liberté des distributions χ^2 ont été sélectionnés de manière à couvrir une gamme importante de degrés de dissymétrie. Les valeurs suivantes de k ont donc été retenues : $1, \ldots, 6, 10$ et 20. Les paramètres des distributions bêta ont été choisis de manière à faire varier le degré d'aplatissement, tout en maintenant la symétrie. Les deux paramètres de forme sont donc identiques ($k_1 = k_2$) et les valeurs suivantes ont été retenues : 0,5, 1, 2, 3, 5 et 10. Les figures 1 et 2 donnent la densité de probabilité de plusieurs de ces distributions, après transformation linéaire. Toutes les distributions n'ont pas été représentées, afin de ne pas surcharger les graphiques. Quant aux mélanges de distributions, nous avons considéré que la moitié des variables sont normales et que les autres ont, à une transformation linéaire près, soit une distribution χ^2 à un degré de liberté soit une distribution uniforme. Dans les deux cas, on a considéré que la différence entre les populations porte soit sur une variable normale (distributions notées $N-C$ et $N-B$), soit sur une variable non normale (distributions notées $C-N$ et $B-N$).

Enfin, trois degrés de recouvrement des populations ont été retenus : 10 %, 20 % et 30 %. Ce degré de recouvrement est déterminé en fonction de la seule variable qui présente une différence de moyenne pour les deux populations. Si X est cette variable, et si μ_1 et $\mu_2 (> \mu_1)$ sont les moyennes de X pour les deux populations, le degré de recouvrement est égal à :

$$[P(X > x_0|A_1) + P(X < x_0|A_2)]/2\,,$$

avec $x_0 = (\mu_1 + \mu_2)/2$, $P(X > x_0|A_1)$ représentant la probabilité que X soit supérieur à x_0, si l'individu provient de la population 1 et $P(X < x_0|A_2)$ étant la probabilité que X soit inférieur à x_0, si l'individu provient de la population 2.

3.3. Estimateurs comparés

Aux 16 estimateurs considérés dans l'étude antérieure (PP, paragraphes 2.3 et 2.4), on a ajouté quatre estimateurs non paramétriques, préférentiellement conçus pour l'estimation du taux d'erreur réel.

Deux de ces estimateurs dérivent directement du taux d'erreur par validation croisée, eCV :

$$eSCV_1 = eCV(n/(n+1)) \quad \text{et} \quad eSCV_2 = eCV(n/(n+3))\,.$$

Ils ont été proposés par Hand [1986], dans le but de réduire la variance de eCV, au détriment d'une augmentation du biais.

L'estimateur $e0$, introduit par Chatterjee et Chatterjee [1983] est une variante du *bootstrap*. Cet estimateur est obtenu pour 100 rééchantillonages, un échantillon de taille n étant prélevé, de manière aléatoire et simple et 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 initial qui ne se retrouvent pas dans l'échantillon bootstrapé. Soit $(e0)_j$ la proportion d'individus mal classés par la règle établie sur la base du $j^{ième}$ échantillon bootstrapé. La valeur de $e0$ est la moyenne arithmétique des 100 proportions d'individus mal classés $(e0)_j$.

Enfin, l'estimateur $e632$ a été proposé par Efron [1983]; il est obtenu par une combinaison linéaire du taux d'erreur apparent eA et du taux d'erreur $e0$ défini ci-dessus :

$$e632 = 0,368\,eA + 0,632\,e0\,.$$

3.4. Critères de comparaison

Au total, 342 situations ont été considérées. Elles résultent de la combinaison de trois valeurs de p, de deux tailles d'échantillons, de 19 types de distributions et de trois taux de recouvrement. Pour chaque situation, on a généré 100 couples d'échantillons. Pour chaque couple, on a déterminé :

- la valeur des 20 estimateurs présentés au paragraphe 3.3;
- la valeur théorique des trois taux d'erreur par la procédure décrite au paragraphe 2. Pour le taux d'erreur optimal et pour le taux d'erreur attendu, les 100 valeurs relatives à une même situation sont identiques, vu la manière dont ils sont déterminés.

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

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

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 et pour chacun des taux théoriques présentés au paragraphe 2. Chacun des 20 estimateurs est donc considéré, dans un premier temps, comme un estimateur du taux optimal, ensuite comme un estimateur du taux réel et, enfin, comme un estimateur du taux attendu.

4. Résultats

4.1. Présentation des résultats

Comme dans l'étude précédente (PP, paragraphe 4.1), nous avons remplacé les erreurs absolues moyennes par des rangs. Pour un taux d'erreur donné et pour une combinaison de facteurs étudiés, le rang de chacun des estimateurs a été déterminé, l'estimateur présentant l'erreur absolue moyenne la plus faible obtenant le rang 1. Ensuite, le rang médian a été calculé pour l'ensemble des 342 situations, mais aussi pour chacune des variantes de chacun des facteurs. Sur la base de ces rangs médians, le classement de chaque estimateur a été déterminé. Ainsi, dans le tableau 1, qui est commenté au paragraphe 4.2, on constate, par exemple, que, globalement l'estimateur eOS est classé en première position, ce qui signifie que la médiane des 342 rangs obtenus par cet estimateur est la plus faible. Par contre, si on ne s'intéresse qu'aux 114 situations pour lesquelles $p = 16$, cet estimateur se situe en cinquième position, la médiane des 114 rangs étant la cinquième plus petite valeur.

Pour quantifier l'importance des différences entre estimateurs, nous avons exprimé, pour chacune des situations, les erreurs absolues moyennes en proportion de l'erreur absolue moyenne de l'estimateur classé en première position. Les résultats ont été exprimés sous la forme d'un graphique en boîtes multiples établi par le logiciel Minitab [Tukey, 1977; X, 1996].

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.

4.2. Estimation du taux d'erreur réel

Le rang de chaque estimateur est donné dans le tableau 1. Dans ce tableau, les estimateurs ont été classés en fonction de leur rang pour l'ensemble des situations. Si on examine le classement global, les deux premières positions sont occupées par eOS et $e632$. A l'autre extrême, on note le très mauvais classement obtenu par eA, eD, ePP, $e3$, $e5$ et $ePPCV$. Le résultat pour ces estimateurs est dans l'ensemble assez stable, quelles que soient la nature des populations, la valeur de p, de p/n et du taux de recouvrement des populations.

TABLEAU 1

Estimation du taux d'erreur réel : rang des estimateurs pour les niveaux des différents facteurs considérés.

	eOS	e632	*eM*	*eL*	*eU*	*eSCV1*	*eSCV2*	*eBoot*	*eCV*	*eJc*	e0	*eO*	*eB*	*eDS*	*ePPCV*	e3	e5	*ePP*	*eD*	*eA*
Ensemble	1	2	3	4	5	6,5	6,5	8	10	10	10	12	13	14	15	16	17,5	17,5	19	20
N	1	5	2	3,5	3,5	6,5	6,5	9,5	11,5	8	9,5	11,5	13	14	15,5	15,5	17	18,5	18,5	20
$C(1)$	2	1	12	16	19	4	5	6	8	11	10	9	3	7	13	15	17	14	18	20
$C(2)$	1,5	1,5	7	13,5	17	12	8	6	13,5	9	10,5	5	3,5	3,5	10,5	15,5	18	15,5	19	20
$C(3)$	1,5	1,5	3	10,5	15	8	7	4,5	12	10,5	13	6	4,5	9	14	16	18	17	19	20
$C(4)$	1	2	3	4	10,5	5,5	5,5	7	8,5	10,5	8,5	12	13	14	15,5	15,5	18	17	19	20
$C(5)$	1	2	3,5	7	14	9,5	6	9,5	12	9,5	13	5	3,5	9,5	15	16	17,5	17,5	19	20
$C(6)$	1,5	1,5	3	6,5	13	9,5	9,5	9,5	13	9,5	13	6,5	4	5	15,5	15,5	17	18	19	20
$C(10)$	1	2	3	6	8	9	4,5	10,5	14	10,5	12	4,5	7	13	15,5	15,5	17,5	17,5	19	20
$C(20)$	1	3	2	4	5	6,5	6,5	10	12,5	11	12,5	8,5	8,5	14	15,5	15,5	17,5	17,5	19	20
$B(.5,.5)$	5	3	4	1	2	8	7	9	11	10	6	12,5	12,5	14,5	14,5	17	18	16	19	20
$B(1,1)$	5	2,5	5	1	2,5	8	7	11	10	9	5	12	13	14	15	16,5	18	16,5	20	19
$B(2,2)$	5,5	3,5	3,5	1	2	7	8	11	10	9	5,5	12	13	14	15	16	18	17	19	20
$B(3,3)$	3	4,5	4,5	1	2	7,5	7,5	9,5	11	9,5	6	12	13	14	15	16	17	18	19,5	19,5
$B(5,5)$	1	5	2,5	4	2,5	6	7	10,5	8,5	10,5	8,5	12	13	14	15	16	17,5	17,5	19	20
$B(10,10)$	1	5	2	3,5	3,5	7	6	10,5	8	10,5	9	12	13	14	15	16	17	18	19	20
$N-C$	1	2	3	4	5	8	7	6	11	9,5	12	9,5	13	14	15	16	17	18	19	20
$N-B$	1	2	3	4,5	4,5	7	6	8	10,5	9	12,5	12,5	10,5	14	16	15	17	18,5	18,5	20
$C-N$	2	1	9,5	16,5	19	7,5	7,5	5	9,5	11,5	11,5	6	3	4	14	13	16,5	15	18	20
$B-N$	5,5	3	5,5	1	2	8	7	9,5	9,5	11	4	12	13	14	15	16	19	17,5	20	17,5
$p/n = .2$	1	2,5	2,5	4	5	8	8	11	11	11	13	8	6	14	15	16,5	18	16,5	19	20
$p/n = .4$	1,5	1,5	3	4	5,5	7,5	5,5	9,5	11	9,5	7,5	12,5	12,5	14	15	16	17,5	17,5	19	20
$p = 4$	1	2,5	4	8	11,5	11,5	9,5	9,5	15,5	13	7	6	2,5	5	15,5	14	17	18	19	20
$p = 8$	1	2	3,5	3,5	5	7	7	9	10,5	10,5	7	12	13	14	15	16	17,5	17,5	19	20
$p = 16$	5	1,5	1,5	3,5	3,5	7	6	10	8,5	8,5	11	12	13	14	15	16	18	17	19	20
$eop = .1$	1	2	3	6	6	11	8	11	13	11	14	9	4	6	15	16	17,5	17,5	19	20
$eop = .2$	1	2	3	4	6,5	6,5	5	10	10	8	10	12	13	14	15	16	17,5	17,5	19	20
$eop = .3$	1	2,5	2,5	4	6	8	8	8	10,5	10,5	5	12	13	14	19	15	16,5	16,5	18	20

Par contre, pour d'autres estimateurs, le classement dépend des situations envisagées. Ainsi, les estimateurs eM, eL et eU sont bien classés pour les populations symétriques, mais leur classement devient très mauvais pour les populations très dissymétriques. Les estimateurs eO, eB et eDS obtiennent un meilleur classement lorsque le nombre de variables est réduit, alors qu'on a la situation inverse pour eM et eU.

Enfin, pour l'ensemble des estimateurs, le facteur qui semble le moins influencer le classement est le taux de recouvrement des populations.

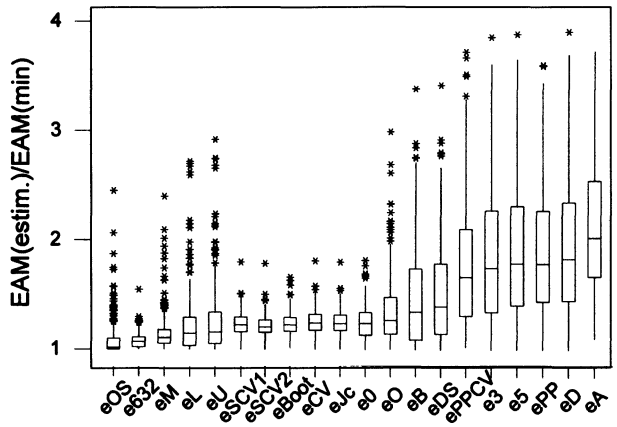

FIGURE 3
Estimation du taux d'erreur réel :
distribution des erreurs absolues moyennes, exprimées en proportion de l'erreur absolue moyenne de la méthode classée en première position

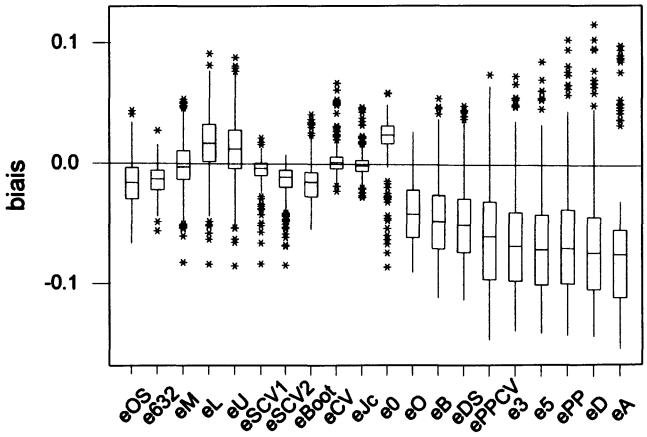

FIGURE 4
Estimation du taux d'erreur réel : distribution des biais

Le graphique des efficacités relatives (figure 3) montre le bon comportement de eOS et $e632$, même si, pour eOS surtout, on observe quelques situations donnant lieu à des erreurs relatives importantes. Pour eM, eL et eU, la dispersion de l'efficacité relative est plus importante, à cause de la mauvaise performance de ces estimations pour les populations très dissymétriques. La plupart des estimateurs basés sur le rééchantillonnage (eCV, $eSCV1$, $eSCV2$, eJC, $eBoot$, $e0$) accusent une perte d'efficacité de 20 à 30 %, assez constante quelle que soit la situation envisagée.

Pour les autres estimateurs, les pertes d'efficacité sont plus importantes et très variables d'une situation à l'autre.

La figure 4 montre que les biais sont, dans l'ensemble, assez faibles pour les estimateurs occupant les 11 premières positions du classement. Pour les autres, le biais est toujours négatif, la médiane dépassant, en valeur absolue, 5 %.

4.3. Estimation du taux d'erreur attendu

Nous ne présentons pas les résultats détaillés pour le taux attendu car les résultats sont tout à fait similaires à ceux obtenus pour le taux réel. En particulier, le classement global est, à une ou deux inversions près dans la partie centrale du classement, identique à celui obtenu pour le taux réel.

4.4. Estimation du taux d'erreur optimal

Le tableau 2 montre que les estimateurs eB et eDS sont globalement les mieux classés, suivis par eO, eOS, $e3$ et $ePPCV$. Les moins bien classés sont $e0$, eU et eL. On note cependant une grande variabilité dans le classement selon la nature de la distribution : les estimateurs eB et eDS ont en effet un classement relativement mauvais pour les distributions bêta caractérisées par un faible coefficient d'aplatissement. Pour ces situations, $e632$, qui globalement est classé en $7^{ième}$ position, offre de bonnes performances.

Les figures 5 et 6 montrent que les différences entre les estimateurs sont nettement moins tranchées pour le taux optimal que pour le taux réel ou attendu. Bien que les meilleurs estimateurs (eB et eDS) présentent un biais faible, ils ont une efficacité relative qui varie fortement avec les situations envisagées. Les résultats sont en fait fortement influencés par quelques situations particulières qui concernent principalement les distributions bêta de paramètres (0,5; 0,5) et (1; 1).

5. Conclusions

Pour le taux d'erreur réel, qui est le taux auquel s'intéresse le plus souvent l'utilisateur, les conclusions sont assez claires. En effet, deux estimateurs, eOS et $e632$, se sont montrés globalement supérieurs aux autres. Ils occupent la première et la seconde position, toutes situations confondues. L'estimateur eOS est un estimateur paramétrique initialement prévu pour l'estimation du taux d'erreur attendu, dans le

TABLEAU 2

Estimation du taux d'erreur optimal : rang des estimateurs pour les niveaux des différents facteurs considérés.

	eB	*eDS*	*eO*	*eOS*	*e3*	*ePPCV*	*e5*	*e632*	*ePP*	*eBoot*	*eD*	*eSCV2*	*eM*	*eSCV1*	*eA*	*eJc*	*eCV*	*eL*	*eU*	*e0*
Ensemble	1	2	3,5	3,5	5,5	5,5	7,5	7,5	9	10	11	12	13	14	15,5	15,5	17	18,5	18,5	20
N	1	2	3	4	5,5	5,5	7	8	9,5	13	9,5	11,5	11,5	15	14	16	18	19	17	20
$C(1)$	5	1,5	9	14,5	3,5	6	3,5	8	7	11	1,5	12	17,5	13	10	14,5	16	19	20	17,5
$C(2)$	1	2	8	11	3	5,5	4	9	7	10	5,5	12	17	13	14	15,5	15,5	19	20	18
$C(3)$	1	2	8	11	3,5	5	3,5	9	6,5	10	6,5	12	17	14	13	15,5	15,5	18,5	20	18,5
$C(4)$	1	2	3	4	7,5	6	9	5	7,5	10	11,5	11,5	15,5	13	17	14	15,5	18	19,5	19,5
$C(5)$	1	2	7	8,5	3	5,5	4	10,5	8,5	10,5	5,5	12	17	14	13	15	16	18	19,5	19,5
$C(6)$	2	1	3,5	7	3,5	5	6	9,5	9,5	11	8	12	14,5	14,5	13	16	17	18	19,5	19,5
$C(10)$	1,5	1,5	4	7	3	5	6	10	8,5	11	8,5	12,5	15	14	12,5	16	17	19	18	20
$C(20)$	2	1	3,5	9	3,5	5	6	10	7,5	11,5	7,5	13	14,5	14,5	11,5	16,5	16,5	19	18	20
$B(.5,.5)$	12,5	14,5	12,5	10	17	14,5	18,5	3	16	11	20	5,5	8,5	5,5	18,5	8,5	5,5	5,5	2	1
$B(1,1)$	12	14	13	2	16,5	15	19	1	18	7	20	6	3,5	8	16,5	9	10	5	3,5	11
$B(2,2)$	5	7,5	6	1	15	12,5	17,5	3	17,5	10	20	9	2	11	19	12,5	14	7,5	4	16
$B(3,3)$	2	3	4	1	13	11	16,5	5,5	16,5	7,5	19	9	5,5	12	18	14	15	10	7,5	20
$B(5,5)$	1	2	3,5	3,5	7	8	9	5	13	12	14	10,5	6	16	17	18	19	15	10,5	20
$B(10,10)$	1	2	3	4	5	6	7	8	10,5	12	9	13	10,5	16	14	17,5	19	17,5	15	20
$N-C$	1	2	3	4,5	4,5	6	7	10	9	11	8	12	13,5	15	13,5	16	17	19	18	20
$N-B$	1	2	3	4	5,5	5,5	7	10	8,5	13	8,5	11	14	15	12	16	17	19	18	20
$C-N$	3	1	10	13,5	3	6	3	9	7	11	5	12	17,5	13,5	8	15	16	19	20	17,5
$B-N$	12	14	13	2	16	15	17,5	1	17,5	8	20	6,5	14	6,5	19	9	10,5	4	4	10,5
$p/n=.2$	1	2	4,5	3	4,5	6	7	9	9	12	9	12	12	14	15	16,5	16,5	18,5	18,5	20
$p/n=.4$	1	2	3,5	3,5	6,5	5	8	6,5	9,5	9,5	11,5	11,5	13	14	15,5	15,5	17	18,5	18,5	20
$p=4$	1,5	1,5	5	7	3	5	5	10	8,5	13,5	8,5	11,5	11,5	15	13,5	16	17	20	18,5	18,5
$p=8$	1	3	3	3	7	5	8	6	9,5	9,5	12	12	12	14	15,5	15,5	17	18,5	18,5	20
$p=16$	3	4	1	2	8,5	7	11	5	10	6	12	8,5	13,5	13,5	17	15	16	18	19	20
$eop=.1$	1,5	1,5	6	9	3,5	3,5	5	11,5	8	10	7	13	14	15	11,5	16	17	19	18	20
$eop=.2$	1,5	1,5	5	3	5	5	7	8	10	11,5	9	11,5	13,5	15	13,5	16	17	18,5	18,5	20
$eop=.3$	2	5	3,5	1	9	13	11,5	3,5	11,5	7	15,5	6	8	10	19	14	15,5	18	20	17

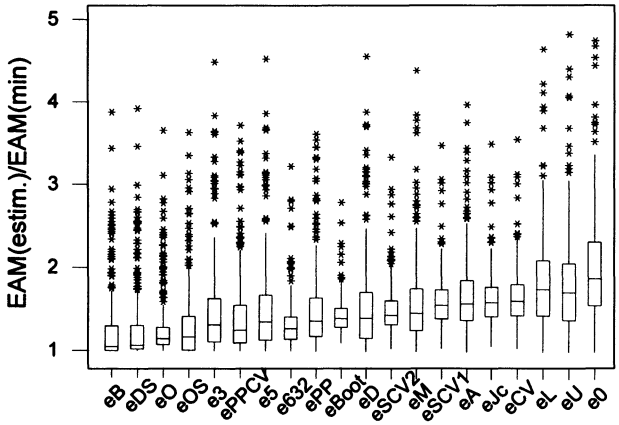

FIGURE 5
Estimation du taux d'erreur optimal : distibution des erreurs absolues moyennes, exprimées en proportion de l'erreur absolue moyenne de la méthode classée en première position

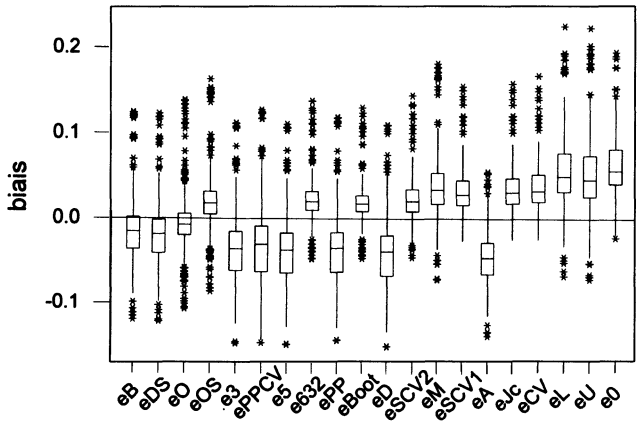

FIGURE 6
Estimation du taux d'erreur optimal : distribution des biais

cas de deux populations normales de même matrice de variances et covariances. $e632$ est un estimateur non paramétrique.

L'estimateur eOS est légèrement supérieur à d'autres estimateurs paramétriques (eM, eL et eU) pour des populations normales ou proches de la normale, mais contrairement à ces derniers, il est assez robuste vis-à-vis de la non normalité, pour autant que celle-ci ne devienne pas excessive. Il est d'ailleurs plus sensible à l'aplatissement des distributions qu'à leur dissymétrie.

L'estimateur $e632$ est supérieur à tous les autres estimateurs non paramétriques. La supériorité de cet estimateur a déjà été mise en évidence par plusieurs auteurs. Des références à ce sujet sont données par Chernick [1999] et par McLachlan [1992]. Pour des distributions normales ou proches de la normale, il est cependant inférieur à eOS; par contre, pour des situations extrêmes (distribution en i, en u, uniforme), il est supérieur à eOS.

Pour le taux d'erreur attendu, les conclusions sont identiques à celles relatives au taux d'erreur réel, le classement des estimateurs étant pratiquement le même que celui obtenu pour le taux réel.

Par contre, pour le taux optimal, l'estimateur eB est le meilleur, sauf pour les situations s'écartant très fortement de la normalité, situations pour lesquelles l'estimateur $e632$ est préférable.

On peut noter aussi que, pour les populations normales, les résultats de la présente étude sont tout à fait concordants avec ceux obtenus lors de l'étude antérieure, ce qui montre bien que les résultats sont reproductibles et non liés au hasard des simulations. Nous avons également déjà signalé que les conclusions rejoignaient celles obtenues par différents auteurs (PP, paragraphe 5).

Rappelons que certaines populations considérées s'écartent très nettement des distributions normales et il est, dès lors, assez naturel que les estimateurs paramétriques qui sont proposés pour des populations normales de même matrice de variances et covariances conduisent, dans les situations extrêmes, à des biais importants. C'est notamment le cas pour les estimateurs eB et eDS considérés comme des estimateurs du taux d'erreur optimal.

Pour ces estimateurs, un biais, positif ou négatif, important s'observe quand, pour la variable présentant une moyenne différente d'une population à l'autre, le taux de recouvrement est très différent de la valeur obtenue par la formule donnant le taux théorique optimal dans le cas de populations normales de même matrice de variances et covariances. Ainsi, par exemple, pour la distribution χ^2 à 1 degré de liberté, le taux de recouvrement de 10 % est obtenu pour une différence de moyennes de 1,41, ce qui donne, par la formule du taux optimal, une valeur égale à :

$$\Phi(-\Delta/2) = \Phi(-0,70) \simeq 0,24\,.$$

Il en résulte que les méthodes d'estimation du taux d'erreur optimal théorique, basées sur l'estimation de Δ, conduisent à une surestimation très importante. Par contre, pour la même distribution mais un taux de recouvrement de 30 %, la différence de moyennes, Δ, est de 1,005 et le résultat de la formule théorique est de 0,307. Dans ce cas, on peut s'attendre à un biais beaucoup plus faible.

On constate cependant que ce n'est que dans des conditions relativement extrêmes que les meilleurs estimateurs paramétriques (eOS pour le taux d'erreur réel et pour le taux d'erreur attendu et eB pour le taux d'erreur optimal) sont surpassés par l'estimateur non paramétrique $e632$.

D'autre part, le fait qu'un même estimateur soit le mieux classé pour l'estimation de plusieurs taux d'erreur signifie que les différences entre les taux théoriques ne peuvent pas être appréciées en pratique, puisqu'une même valeur est utilisée comme estimation de deux ou trois valeurs différentes. Ainsi, par exemple, la distinction entre

taux d'erreur réel et taux d'erreur attendu n'a pas d'intérêt si ces taux doivent être estimés. De même pour des populations nettement non normales, la distinction, sur la base d'estimation, des trois taux d'erreur définis au paragraphe 2 n'est pas possible, l'estimateur $e632$ étant préconisé pour l'estimation des trois taux.

Bibliographie

CHATTERJEE S., CHATTERJEE S. [1983]. Estimation of misclassification probabilities by bootstrap methods. *Commun. Stat. Simul. Comput.* 12, 645-656.

CHERNICK M.R. [1999]. *Bootstrap methods : a practioner's guide*. New York, Wiley, 264 p.

EFRON B. [1983]. Estimating the error of a prediction rule : improvement on cross-validation. *J. Amer. Stat. Assoc.* 78, 316-331.

HAND D.J. [1986]. *Cross-validation in error rate estimation*. Proc. XIIIth Inter. Biometric Conference. Alexandria, Virginia, Biometric Society, 13 p.

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

PIRAUX F., PALM R. [2001]. Etude empirique des estimateurs des taux d'erreur en analyse discriminante. *Rev. Stat. Appl.* 49 (4), 71-85.

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

X [1996]. *Minitab reference manual : release 11 for Windows.* P.A. State College, Minitab, 1.052 p.