Comput. Math. Applic. Vol. 15, No. 1, pp. 29–37, 1988
Printed in Great Britain. All rights reserved

0097-4943/88 $3.00 + 0.00

ESTIMATION OF ERROR RATE FOR LINEAR DISCRIMINANT FUNCTIONS BY RESAMPLING: NON-GAUSSIAN POPULATIONS

M. R. Chernick,[1] V. K. Murthy[2] and C. D. Nealy[2]

[1]The Aerospace Corporation, P.O. Box 92957, M3-644, Los Angeles, CA 90009, U.S.A.

[2]Image Processing Laboratory, Hughes Aircraft Company, El Segundo, CA 90245, U.S.A.

(*Received* 8 *April* 1987)

Communicated by E. Y. Rodin

Abstract—This article presents simulation results comparing various resampling estimators of classification error rate for linear discriminant type classification algorithms. Three non-Gaussian multivariate populations are studied namely, exponential, Cauchy and uniform. Simulations are conducted for small sample sizes, two-class and three-class problems and 2-D, 3-D and 5-D distributions. Estimation procedures and sample sizes are the same as in our previous study of Gaussian populations; again 200 bootstrap replications are used for each simulation trial. For exponential and uniform distributions the 0.632 estimator generally performs best. However, for Cauchy distributions the convex bootstrap and the e_0 often outperform the 0.632 estimator.

1. INTRODUCTION

This paper deals with the estimation of classification error rates for linear discriminant rules when the unknown population distributions have multivariate distributions which are not Gaussian. The basic problem is that the naive estimator (referred to in the literature as the apparent error rate or the resubstitution estimate) is known to be optimistically biased. This is particularly true when the number of training samples is small and the true error rate is high. The apparent error rate simply estimates the misclassification probability by applying the estimated discriminant rule to classify the training set. The estimate is then the number of misclassifications divided by the number of training vectors.

There has been a great deal of research in the estimation of misclassification probability, as can be seen from the extensive bibliography by Toussaint [1]. In recent years, interest has been renewed due to the work of Glick [2] and Efron [3]. We simulate bivariate exponential, bivariate uniform and 2-D, 3-D and 5-D Cauchy populations. This choice was made to learn something about the effect of skewness and distribution tail length on the estimates. The approach is the same as in our study for Gaussian populations [4]. We consider the same seven estimators (the apparent error rate and six resampling methods).

The method of cross-validation (also referred to in the literature as the "U method" or "leave-one-out" estimator) was first suggested by Lachenbruch [5] and popularized by Lachenbruch and Mickey [6]. This procedure removes most of the bias of the apparent error rate by computing the n estimates (n is the number of training vectors) of the discriminant functions obtained by leaving out one training vector each time. The discriminant functions are then used to classify the training vectors left out. The estimate is obtained by counting the number misclassified and dividing by n.

Unfortunately, this estimator has been found to have a large variance [2–4]. Alternatives to cross-validation were proposed by Glick [2] and have been modified and studied by Snapinn and Knoke [7, 8]. Efron [3] proposed resampling or bootstrap-type procedures as an alternative. He demonstrated improvement through a simulation of a few Gaussian cases using a small number of training vectors.

Most promising of the bootstrap-type estimates was the 0.632 estimator. Efron [3] proposed the 0.632 estimator and found it to be better than the other resampling estimators. The 0.632 estimator is formed by a weighted average of the apparent error rate and the e_0 estimator with weights 0.368 and 0.632, respectively. The e_0 estimate is obtained by simply totaling the number of training vectors misclassified among those training vectors not included in the bootstrap samples and dividing by the total number of training vectors left out of the bootstrap samples.

Subsequent simulation studies [4, 9] confirm the superiority of the 0.632 estimator. In particular, recent results by Snapinn and Knoke [9] indicate that it is competitive with other estimators for non-Gaussian populations. We shall show why the 0.632 estimator works so well and shall demonstrate through simulations that for certain Cauchy populations it may not be the best estimator. A more detailed study of the sensitivity of the 0.632 estimator to distribution tail length is the subject of a future paper of the authors using Pearson VII distributions in a systematic approach to multivariate distribution generation, as proposed by Johnson *et al.* [10, 11].

Modifications of some of the current estimators may produce improvement and these modifications are proposed for future comparison with the smoothed estimators of Snapinn and Knoke [9]. Knoke [12] gives an interesting summary of error rate estimation and the value of bootstrap methods is emphasized by McLachlan [13].

2. TECHNICAL APPROACH

In this paper the 0.632, bootstrap, convex bootstrap, e_0, apparent error rate and MC estimators are computed for each of 200 simulations for a variety of true error rates and multivariate distributions. The estimators are the same as those defined and compared by Chernick *et al.* [4]. The true expected unconditional error rate is used instead of the Mahalanobis distance as a measure of population separation. It seems to us to be a more natural parameter and is interpretable for non-Gaussian populations, whereas the Mahalanobis distance is a natural measure for Gaussian populations but is not easily interpretable for non-Gaussian populations.

We shall now describe the various populations studied and then give the definitions of the estimators. It should be pointed out, as has been noted by Sorum [14], that there are at least three misclassification probabilities that can be estimated. Page [15] shows that the choice of estimator depends on the type of error rate of interest and sample size. Here, as in our previous work, we estimate the expected probability of misclassification given a fixed training set. This is replicated for 200 training sets to obtain our measure of estimator performance, the unconditional mean square error, as in Refs [3, 4, 7–9]. We also average the 200 expected probabilities of misclassification to obtain our approximate "true error rate", our measure of population separation.

We simulated *bivariate exponential distributions* of the type described by Gumbel [16]. The general procedure for generating bivariate exponential and uniform random vectors of this type is described by Chernick [17] (see also Ref. [11]), and is defined as follows:

$$F(x_1, x_2) = F_1(x_1)\, F_2(x_2)\{1 + A[1 - F_1(x_1)][1 - F_2(x_2)]\}, \tag{1}$$

where F_1 and F_2 are the univariate exponential or uniform cumulative distributions and $|A| \leqslant 1$ ($A = 0.5$ in the simulations reported). Values of A other than 0.5 were tried but did not appear to affect the results.

For the Cauchy distributions the following autoregressive scheme is used to generate 2-D, 3-D and 5-D distributions, with m, a constant location parameter:

$$\begin{aligned} X_1 &= \rho X_{i-1} + \epsilon_i + (1 - \rho)\mathrm{m} \\ X_1 &= \mathrm{m} + C_1, \end{aligned} \tag{2}$$

where C_1 is a Cauchy random variable with scale parameter 1 and location parameter 0, ϵ_i are independent and Cauchy with scale parameter $1 - |\rho|$ and location parameter 0 and $|\rho| \leqslant 1$ ($\rho = 0.5$ in the simulations reported). Results for other values of ρ were computed but there appeared to be no effect due to the autoregressive parameter.

For the exponential and uniform cases, simulations are run for *true error rates* ranging from 0.1 to 0.5, for two-class problems with 2-D feature vectors and training sample sizes of 14, 20 and 29. For the uniform case three-class problems are also considered with true error rates ranging from 0.05 to 0.67 with training sample sizes of 20 and 29.

For the Cauchy case two-class and three-class problems are considered for 2-D, 3-D and 5-D *feature vectors*. For the two-class and 2-D problems, training sample sizes are 14, 20 and 29, and only 20 and 29 for the three-class or 3-D and 5-D problems. Error rates vary from 0.05 to 0.5 for two-class problems and from 0.05 to 0.67 for three-class problems.

For the Cauchy and uniform distributions, different true error rates are obtained by varying the shift in location parameters from one population to the next. For the bivariate exponential distribution the cumulative distributions F_1 and F_2 in equation (1) are given by

$$F_i(x) = 1 - \exp(-\lambda x), \quad i = 1,2, \tag{3}$$

where λ is the rate parameter. Since the mean and variance both depend on λ, separation of the populations is achieved by varying λ. We did not consider any two-parameter exponential distributions.

The apparent error rate and the cross-validation (U method) estimates are determined as we have previously defined them. The standard bootstrap estimated is obtained through the generation of bootstrap samples. A bootstrap sample is obtained by sampling with replacement from the empirical distribution for the training set. Several bootstrap samples are generated (in our simulations we generate 200 bootstrap samples for each of 200 simulation trials).

The bias of the apparent error rate is estimated by the bootstrap sample analog to equation (2.10) of Efron [3, p. 317]. This estimate of bias is added to the apparent error rate to obtain the bootstrap estimate of misclassification probability. We call this procedure the "standard bootstrap" to distinguish it from other variants of the bootstrap.

Chernick and Murthy [18] investigated properties of bootstrap samples based on a connection between bootstrap sampling and the classical occupancy problem. In a bootstrap sample the probability is approx. 0.368 that an observation vector will *not* be included in a bootstrap sample. The exact probability is a function of sample size. In small samples this probability is <0.368 (e.g. for $n = 14$, $p = 0.354$ and for $n = 2$, $p = 0.250$).

The MC estimator works exactly like the bootstrap except that individual "bootstrap" samples are controlled to leave out a fixed proportion of the training set and to include fixed proportions once, twice and three times. These proportions are based on the asymptotic repetition frequencies for bootstrap samples as given in Chernick and Murthy [18].

When the feature vectors come from an absolutely continuous distribution such as the multivariate Gaussian or the multivariate exponential, uniform and Cauchy distributions considered in this paper a problem arises with the standard bootstrap in small samples, namely that the empirical distribution is discrete and thus is only a rough approximation to the population distribution. The convex bootstrap overcomes this problem by taking a convex combination of two independent bootstrap training vectors. This allows for observations to be chosen in between training samples. The choice of independent bootstrap training vectors was done for simplicity but it does lead to inconsistent estimates.

The e_0 estimate, was first defined by Efron [3] and used by him in the calculation of the 0.632 estimator. However, Efron did not compare e_0 with the other estimates in his simulations, but Chernick *et al.* [4] did. Chatterjee and Chatterjee [19] proposed the use of e_0 as a "modified bootstrap" approach but did not explicitly refer to their estimate as e_0.

3. SIMULATION RESULTS

Tables 1–3 summarize the simulation results for the Cauchy, uniform and exponential cases, respectively. Listed is the number of cases for which the various estimators ranked first, second and third based on unconditional mean square error. A case is defined by a particular number of dimensions, classes, training samples and true error rate. Each case is based on 200 simulation trials. We use the notation (2,2,14) to denote a two-class 2-D problem with 14 training vectors. The estimators 0.632, MC, e_0, standard bootstrap, convex bootstrap, cross-validation and apparent error rates are denoted by 632, MC, E0, BOOT, CONV, U and APP, respectively. So, for example,

Table 1. Performance summary—Cauchy case

	632	E0	MC	BOOT	CONV	U	APP	TOTAL
1st	29	47	9	26	36	0	5	152
2nd	45	12	26	46	21	1	1	152
3rd	18	3	46	22	50	7	6	152
Total	92	62	81	94	107	8	12	456

Table 2. Performance summary—uniform case

	632	E0	MC	BOOT	CONV	U	APP	TOTAL
1st	54	18	1	1	1	0	7	82
2nd	17	33	9	11	4	2	6	82
3rd	6	4	14	28	20	4	6	82
Total	77	55	24	40	25	6	19	246

Table 3. Performance summary—exponential case

	632	E0	MC	BOOT	CONV	U	APP	TOTAL
1st	24	5	7	4	0	0	8	48
2nd	6	4	16	17	3	0	2	48
3rd	13	1	8	15	7	0	4	48
Total	43	10	31	36	10	0	14	144

with the exponential distribution there are 48 cases corresponding to the various true error rates, dimensions, classes and training sample sizes.

The results indicate that the 0.632 estimator is superior in the exponential and uniform cases. However, in the Cauchy case the e_0, bootstrap and convex bootstrap rank first more often than 0.632, and the bootstrap and convex bootstrap rank among the top three more often than 0.632.

Results for the individual cases are shown in Figs 1–8. Each figure contains four graphs. The unconditional root mean square error (r.m.s.) for the 0.632, e_0, MC and bootstrap estimates are plotted on one graph and for the U method, convex bootstrap and apparent error rate on another. Similarly, the bias is plotted on the other two graphs. Other cases included in the summary table were deleted from the plots for simplicity since results for those cases were similar to Figs 1–8.

General trends to be noticed are that for true error rates between 0.1 and 0.4 the 0.632 estimator has the lowest mean square error for two-class problems. For three-class problems the 0.632 estimator is best when the error rate is between 0.1 and 0.5. When the error rate is >0.4 for two-class problems and >0.5 for three-class problems the e_0 estimate is best. For error rates <0.1 all estimates are equally good and even the apparent error rate sometimes has the lowest mean square error. These results are similar to our results for the Gaussian case [20].

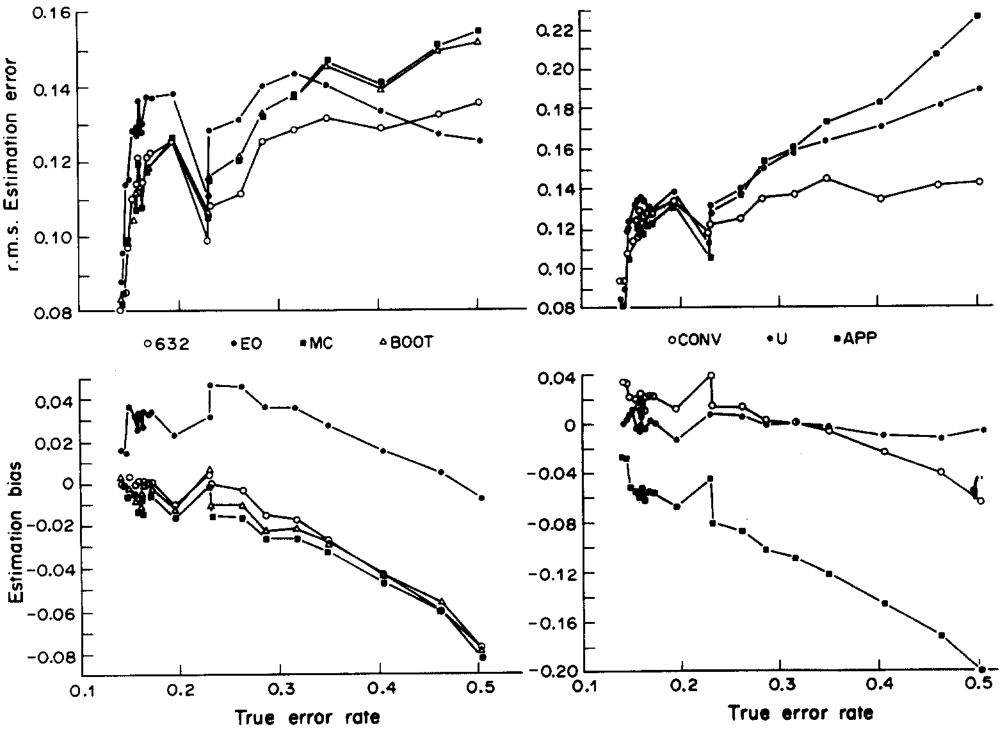

Fig. 1. Exponential (2,2,14).

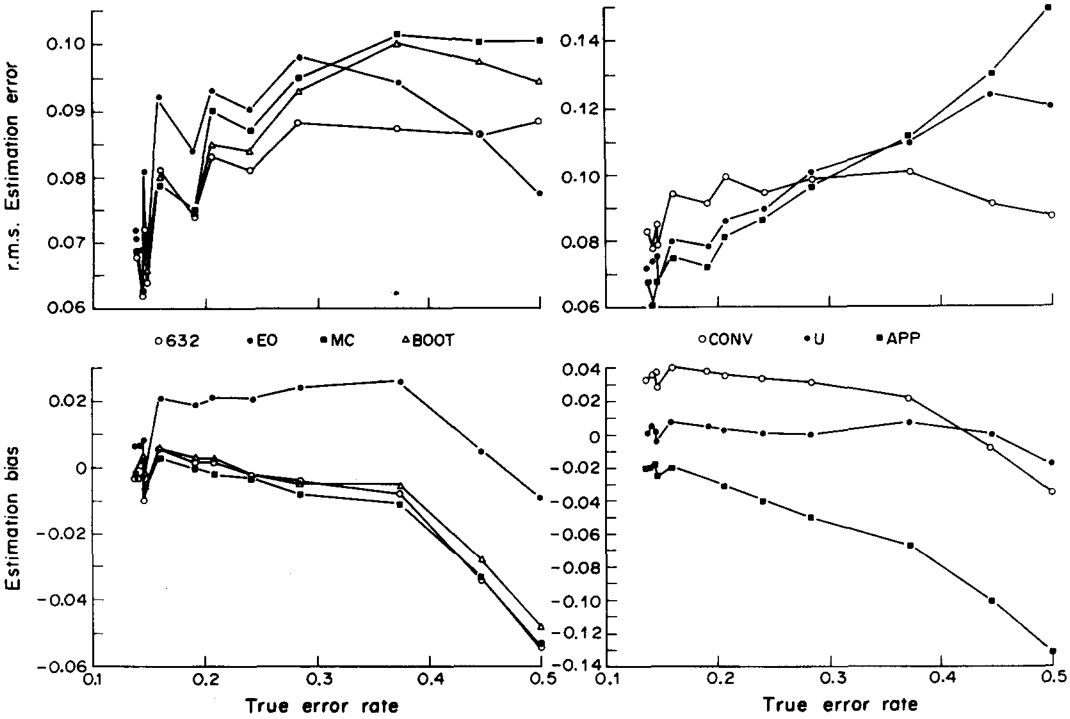

Fig. 2. Exponential (2,2,29).

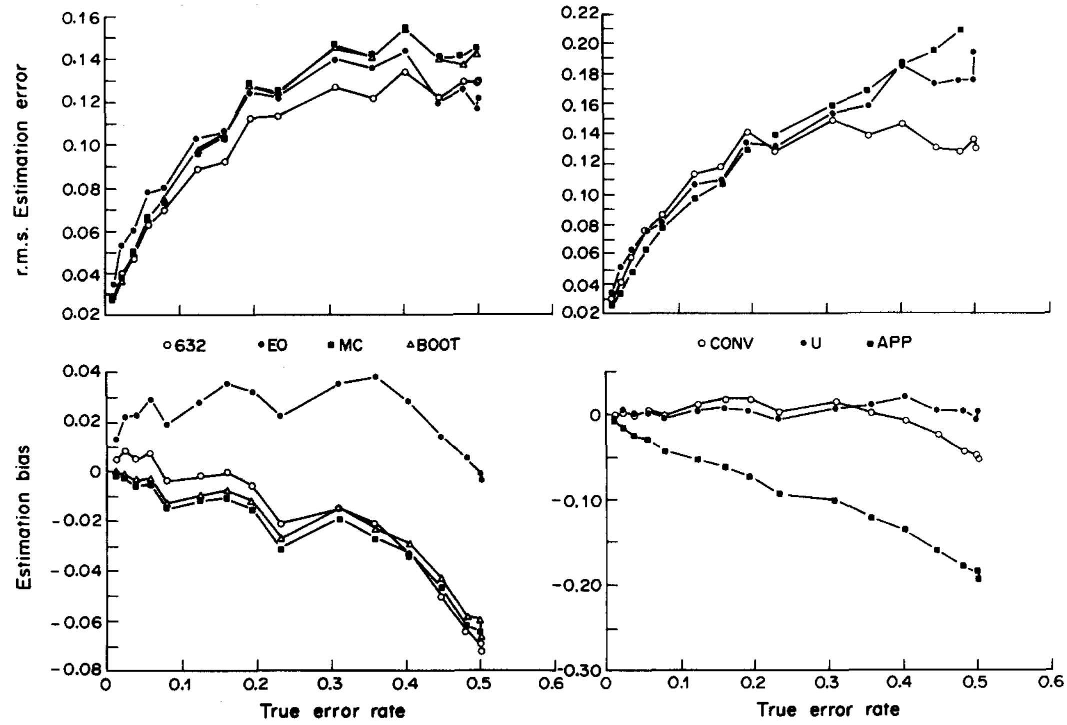

Fig. 3. Uniform (2,2,14).

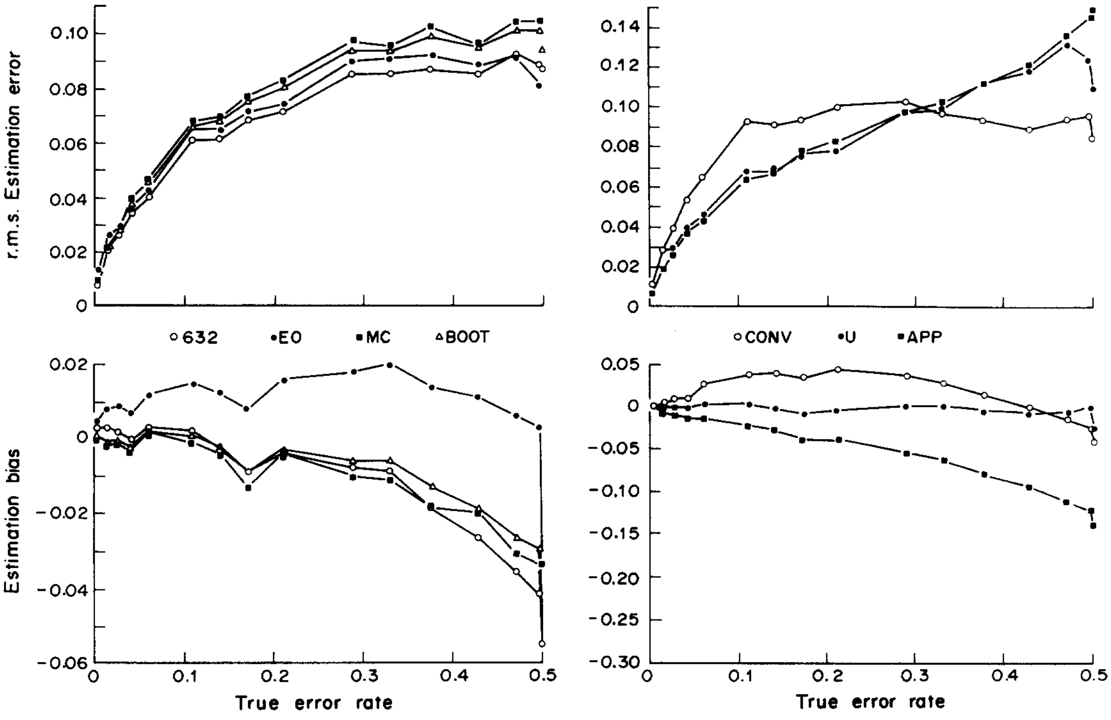

Fig. 4. Uniform (2,2,29).

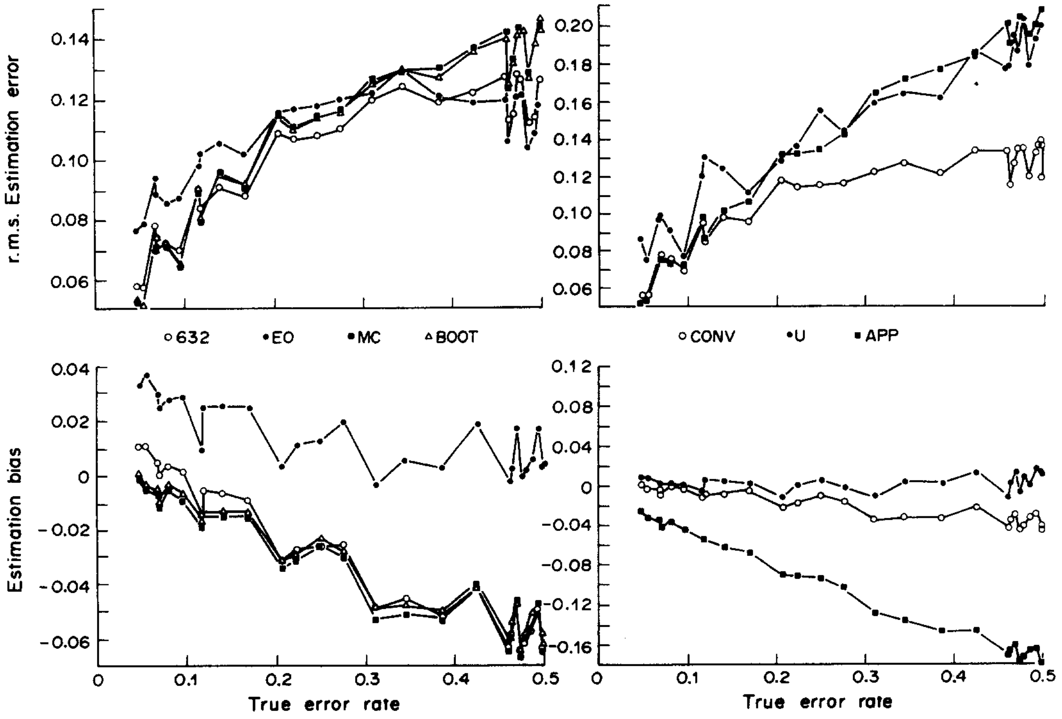

Fig. 5. Cauchy (2,2,14).

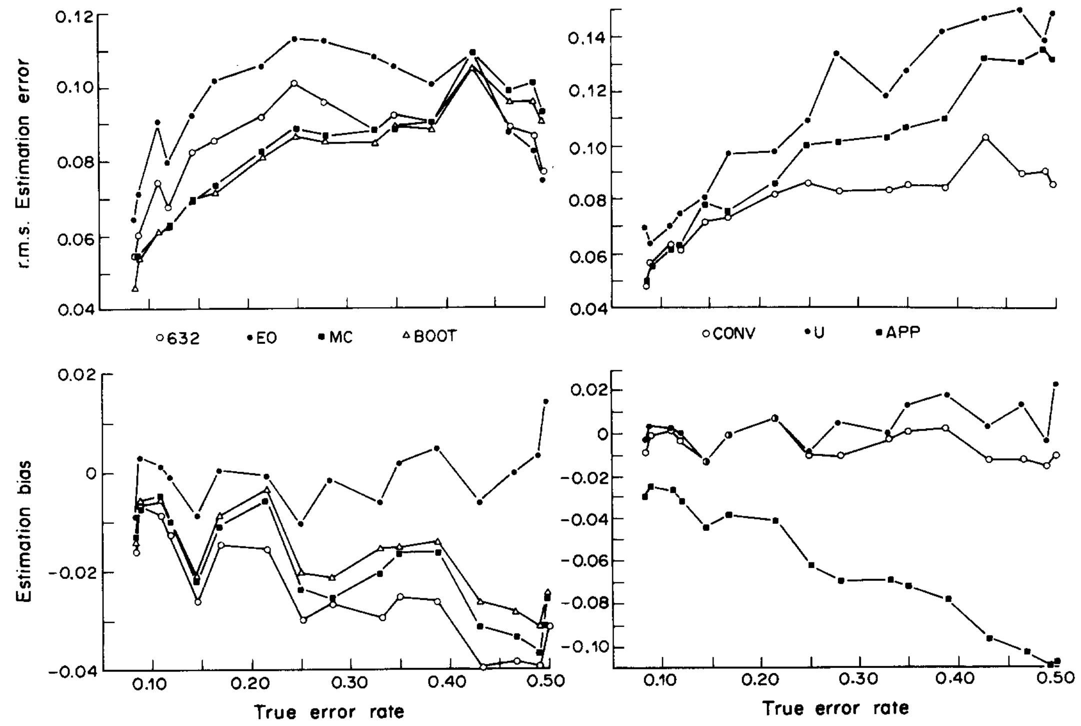

Fig. 6. Cauchy (2,2,29).

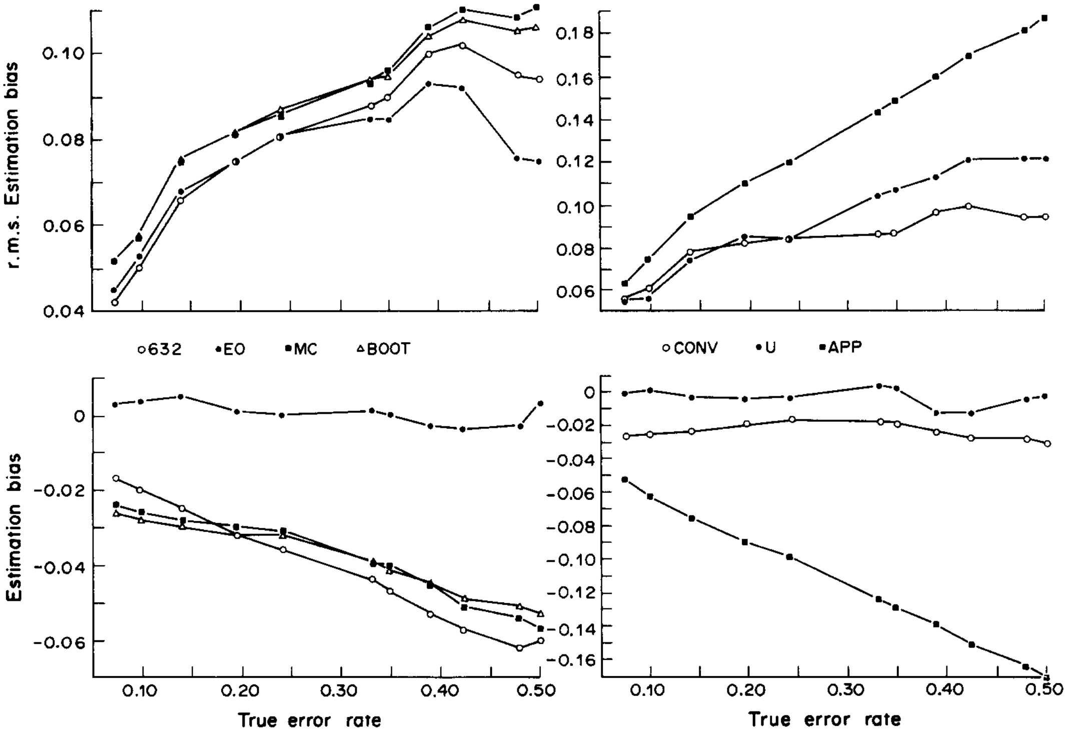

Fig. 7. Cauchy (2,5,29).

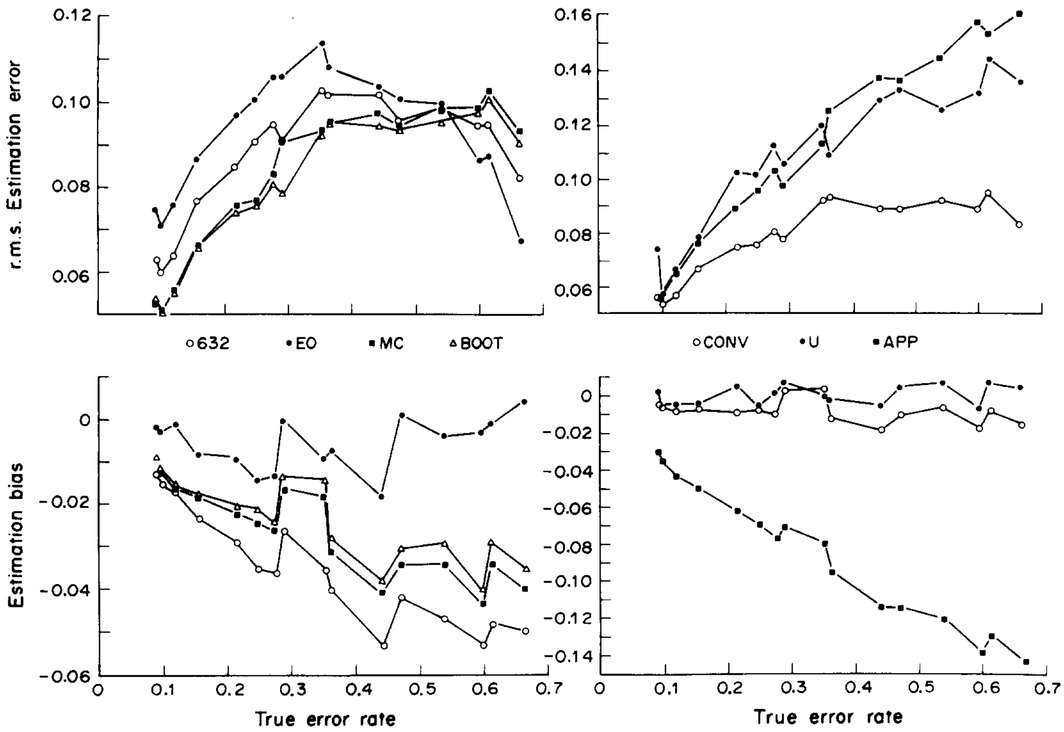

Fig. 8. Cauchy (3,2,29).

Also, similar to the Gaussian results are the trends in the bias for the apparent error rate and the e_0 estimate. The apparent error rate has a negative bias which decreases in a nearly linear fashion with increasing true error rate. The e_0 estimate has a positive and decreasing bias with increasing error rate.

These trends in the bias of the apparent error rate and e_0 estimate partially explain the general success of the 0.632 estimator. The 0.632 estimator appropriately weights two estimates with small variance and opposite biases. However, a scheme which adaptively weights the two estimates depending on an unbiased estimate of the true error rate such as the U estimator might do better at the high error rates since more weight should be given to e_0 there. Also, at the low error rates more weight given to the apparent error rate might lead to improvement.

There is one exception to these trends in bias. For the Cauchy case when the sample sizes are 20 or 29 the e_0 estimator does not have the usual positive bias. Thus, the near cancellation of biases does not occur and the 0.632 estimator is no longer superior to the others. Apparently, the heavy-tailed behaviour of the Cauchy samples has an effect on the bias of e_0. This effect does not show up when $n = 14$. This result is not peculiar to this particular multivariate Cauchy distribution. The authors in a forthcoming paper have obtained similar results for heavy-tailed Pearson VII bivariate distributions including a bivariate Cauchy distribution.

4. SUMMARY AND CONCLUSIONS

The MC estimator, which was expected to be an improvement on the bootstrap, appears to perform similarly to the standard bootstrap. For the non-Gaussian cases studied here, performance results are very similar to previous results for Gaussian cases with the exception of the heavy-tailed Cauchy distribution.

Consequently, we conclude that relative performance may not depend much on skewness but may change for heavy-tailed distributions. If the expected error rate is $< 10\%$ all estimators perform well including the apparent error rate. The 0.632 estimator is generally best when the error rate is between 0.10 and 0.40 for two-class problems and between 0.10 and 0.50 for three-class problems. The e_0 estimate is best when the error rate is > 0.40 for two-class problems and > 0.50

for three-class problems. For the Cauchy case when the training sample size is 20 or 29 the convex bootstrap often performs the best at the low error rates and the e_0 at the high error rates.

A complication that overcomes the consistency problem for the convex bootstrap would be to take convex combinations of neighboring training vectors or to choose observations in an appropriately chosen small neighborhood of a bootstrap observation. This does however increase the computational complexity of an already computer-intensive estimation procedure. Nevertheless, such a modification may be worth considering in the future.

Smoothing of the empirical distribution should help in small sample size problems and our proposed modification to the convex bootstrap may be promising particularly in heavy-tailed situations such as the Cauchy. Another approach to smoothing the estimates is provided by the NS estimator of Snapinn and Knoke [8]. Snapinn and Knoke [9] suggest a bootstrap-type adjustment to the NS estimator that looks promising. The most promising of these estimators should be compared with our proposed modifications to 0.632 and the convex bootstrap over a variety of populations, particularly populations with heavy tails. In addition an alternative approach would be to estimate the tails of a multivariate distribution as for example by the maximum-entropy histogram estimation procedure.

Acknowledgement—We wish to thank J. Knoke for discussions regarding current research and for sending us a preprint of his latest paper with S. Snapinn.

REFERENCES

1. G. T. Toussaint, Bibliography on estimation of misclassification. *IEEE Trans. Inf. Theory* **IT-20,** 472–479 (1974).
2. N. Glick, Additive estimators for probabilities of correct classification. *Pattern Recogn* **10,** 211–222 (1978).
3. B. Efron, Estimating the error rate of a prediction rule: improvements on cross validation. *J. Am. statist. Ass.* **78,** 316–331 (1983).
4. M. R. Chernick, V. K. Murthy and C. D. Nealy, Application of bootstrap and other resampling techniques: evaluation of classifier performance. *Pattern Recogn Lett.* **3,** 167–178 (1985).
5. P. A. Lachenbruch, An almost unbiased method of obtaining confidence intervals for the probability of misclassification in discriminant analysis. *Biometrics* **23,** 639–645 (1967).
6. P. A. Lachenbruch and M. R. Mickey, Estimation of error rates in discriminant analysis. *Technometrics* **10,** 1–11 (1968).
7. S. M. Snapinn and J. D. Knoke, Classification error rate estimators evaluated by unconditional mean squared error. *Technometrics.* **26,** 371–378 (1984).
8. S. M. Snapinn and J. D. Knoke, An evaluation of smoothed classification error-rate estimators. *Technometrics* **27,** 199–206 (1985).
9. S. M. Snapinn and J. D. Knoke, Improved classification error rate estimation: bootstrap or smooth? Unpublished manuscript (1986).
10. M. E. Johnson, *Multivariate Statistical Simulation.* Wiley, New York (1987).
11. M. E. Johnson, C. Wang and J. S. Ramberg, Generation of continuous multivariate distributions for statistical applications. *Am. J. math. Mgmt Sci.* **4,** 225–248 (1984).
12. J. D. Knoke, The robust estimation of classification error rates. *Comput. Math. Applic.* **12A,** 253–260 (1986).
13. G. J. McLachlan, Assessing the performance of an allocation rule. *Comput. Math. Applic.* **12A,** 261–272 (1986).
14. M. Sorum, Three probabilities of misclassification. *Technometrics* **14,** 309–316 (1972).
15. J. T. Page, Error-rate estimation in discriminant analysis. *Technometrics* **27,** 189–198 (1985).
16. E. Gumbel, Bivariate exponential distributions. *J. Am. statist. Ass.* **55,** 698–707 (1960).
17. M. R. Chernick, Generating bivariate distributions with specified marginal distributions. Unpublished manuscript (1986).
18. M. R. Chernick and V. K. Murthy, Properties of bootstrap samples. *Am. J. math. Mgmt Sci.* **5,** 161–170 (1985).
19. S. Chatterjee and S. Chatterjee, Estimation of misclassification probabilities by bootstrap methods. *Communs Statist. Simuln Computn* **12,** 645–656 (1983).
20. M. R. Chernick, V. K. Murthy and C. D. Nealy, Correction note to application of bootstrap and other resampling techniques: evaluation of classifier performance. *Pattern Recogn Lett.* **4,** 133–142 (1986).