Nonparametric rank discrimination method

Jan Ćwik and Jan Mielniczuk
Polish Academy of Sciences, Warsaw, Poland

Abstract: In the paper a new rank approach to solve the nonparametric discriminant problem is presented. The method is based on comparison of kernel density estimators constructed for samples of ranks. The proposed bandwidth choice is motivated by an asymptotic representation for a kernel density estimator in this case. The investigated approach is compared with Rank Linear Discriminant Function (RLDF) and Rank Quadratic Discriminant Function (RQDF) methods. It is indicated that the RLDF and RQDF methods have larger mean probability of error than the proposed methods when normality assumptions are violated.

Keywords: Rank discrimination method; Data-dependent bandwidth; Kernel density estimator; Rank linear (quadratic) discriminant function; Mean integrated squared error; Smoothing

1. Introduction

In the paper we consider an approach to nonparametric discriminant problem when a discriminant function depends only on ranks of observations in a pooled training sample. The use of such methods was advocated by several authors, e.g. Randles at al. (1978), Conover and Iman (1980) and Broffitt (1982). One of the main advantages of using such methods is their invariance with respect to increasing transformation of underlying distributions. For a discriminant problem, it means that when underlying probability measures P_1^T, P_2^T are transformed measures of P_1, P_2 for some increasing transformation T, properties of rank discriminant method for the pair (P_1^T, P_2^T) are the same as for the pair (P_1, P_2). It follows that when P_1, P_2 are known but T is unknown, properties of a rank discriminant method may be evaluated (e.g. using computer intensive methods) for the pair (P_1, P_2) only. Moreover, robustness to outliers is listed as the advantage of rank methods in discrimination. The use of the Rank Linear Discriminant Function (RLDF) and Rank Quadratic Discriminant Function (RQDF) was proposed by Conover and Iman (1980); for the review of this approach see Broffitt (1982).

Further development of this methodology was hindered by the fact that, due to dependencies between ranks, theoretical properties of classical discriminant

SSDI 0167-9473(93)E0043-4

functions applied to rank data were unknown. Thus, a sound application of these functions was, in fact, impossible. This in particular applies to empirical Bayes rules based on estimation of densities, since the problem of choosing an appropriate smoothing parameter is essential for the performance of such rules (see Hand (1982) for the review of such methods).

In the paper, we state (Section 3) an asymptotic representation for a kernel estimator of a density based on ranks. The result suggests that the mean integrated squared error of kernel estimator in the considered case differs from that obtained for an i.i.d. sample and indicates that direct application of bandwidth choice procedures developed for the i.i.d. case is not methodologically valid. Moreover, the result yields (Section 4) some theoretical justification for corrected bandwidth choice for nonparametric rank discrimination based on kernel estimation of densities. Performance of a resulting discriminant rule is investigated in Section 5 for various pairs of univariate distribution functions (F_1, F_2) and compared with performance of RLDF and RQDF. We show that in several situations, when normality assumption is violated, the proposed methods yield smaller mean probability of error than RLDF and RQDF. On the other hand, RLDF and RQDF exhibit generally smaller variability than the introduced methods.

It should be stressed that the discussed method is of limited applicability since it covers only the one dimensional case. However, it could be applied in situations when the original multivariate data is transformed to one dimension (by some 'optimal' transformation) and discriminant problem has to be solved using the resulting data.

2. Definitions

Let F_1 and F_2 be univariate distribution functions and f_1, f_2 corresponding densities with respect to Lebesgue measure on $\mathbb{R}^1$. Assume that F_1, F_2 describe the distribution of some quantity in two populations. Let Z be a random variable with distribution $\pi F_1 + (1-\pi)F_2$, where $1 > \pi > 0$ and π is known. Bayes discriminant rule classifying Z to the first (1) or to the second (2) population is defined as

$$I_{(F_1, F_2)}(Z) = \begin{cases} 1, & \text{if } \pi f_1(Z) > (1-\pi) f_2(Z); \\ 2, & \text{otherwise}. \end{cases} \tag{2.1}$$

(cf e.g. Devroye and Györfi (1985), Chapter 10).

Denote by $X_1, \dots, X_m$, $Y_1, \dots, Y_n$ two independent i.i.d. samples pertaining to F_1 and F_2, respectively, and by $R_{1i}(R_{2j})$ the ranks of $X_i(Y_j)$ in the pooled sample $X_1, \dots, X_m$, $Y_1, \dots, Y_n$:

$$R_{1i} = \#\{k\colon X_k \le X_i\} + \#\{k\colon Y_k \le X_i\},$$

$$R_{2j} = \#\{k\colon X_k \le Y_j\} + \#\{k\colon Y_k \le Y_j\}.$$

We assume that m and $n = n(m)$ are deterministic sample sizes such that $m/(m+n(m))$ tends to π when $m \to \infty$. We seek an empirical analogue $\hat{I}$ of I based on ranks R_{1i}, R_{2j} and the rank of Z in the pooled sample. To this end put $H(x) = \pi F_1(x) + (1-\pi)F_2(x)$ and observe that

$$I_{(F_1,\, F_2)}(Z) = I_{(F_1^H,\, F_2^H)}(H(Z)), \tag{2.2}$$

where F_i^H, $i = 1, 2$ denotes the distribution function of $H(X_1)$ and $H(Y_1)$, respectively. Let $\bar{f}_i$ $(i = 1, 2)$ denote the density of $F_i^H(i = 1, 2)$ with respect to Lebesgue measure; observe that the densities exist and $\pi\bar{f}_1(x) + (1-\pi)\bar{f}_2(x) = 1$. Moreover, put $F_{1m}(F_{2n})$ for the empirical distribution function based on the first (the second) sample. The equality (2.2) yields in a natural way a desired discriminant rule based on ranks

$$\hat{I} = \begin{cases} 1, & \text{if } \pi\hat{\bar{f}}_1(H_N(Z)) > (1-\pi)\hat{\bar{f}}_2(H_N(Z)); \\ 2, & \text{otherwise,} \end{cases} \tag{2.3}$$

where $H_N(x) = (m/N)F_{1m}(x) + (n/N)F_{2n}(x)$, $N = m+n$ and $\hat{\bar{f}}_1(\hat{\bar{f}}_2)$ is some estimator of $\bar{f}_1$ $(\bar{f}_2)$ based on $R_{11}, \ldots, R_{1m}$ $(R_{21}, \ldots, R_{2n})$. In what follows we consider a kernel estimator of $\bar{f}_1$ of the following form

$$\hat{\bar{f}}_1(x) = \frac{1}{mb}\sum_{i=1}^{m}\left\{K\left(\frac{x - H_N(X_i)}{b}\right) + K\left(\frac{x + H_N(X_i)}{b}\right)\right.$$
$$\left. + K\left(\frac{x + H_N(X_i) - 2}{b}\right)\right\}, \tag{2.4}$$

and $\hat{\bar{f}}_2$ is defined analogously, with $\{H_N(Y_j)\}_{j=1}^{n}$ replacing $\{H_N(X_i)\}_{i=1}^{m}$. In the above definition, K is an arbitrary probability density function (kernel) and $b = b_m > 0$ is a smoothing parameter, possibly depending on the underlying sample. Observe that $H_N(X_i)$ is equal to R_{1i}/N. Note also that (2.4) is the ordinary kernel estimator with modification near boundaries (cf e.g. Schuster (1985)) constructed for the sample $\{H_N(X_i)\}$. For the other methods to deal with boundary effects in case of the density estimation see e.g. Gasser et al. (1985) and Diggle and Marron (1988). Estimator (2.4) was introduced in Behnen et al. (1983) as a tool to construct asymptotically optimal rank statistic for two-sample problem. In their paper uniform strong consistency of (2.4) is proved.

More natural alternative to the decision rule (2.3) is the following rule

$$\tilde{I}_{12} = \begin{cases} 1, & \text{if } \pi\hat{\bar{f}}_1(H_N(Z_i)) > \tfrac{1}{2}; \\ 2, & \text{otherwise,} \end{cases} \tag{2.5}$$

which uses the known relationship between $\bar{f}_1$ and $\bar{f}_2$. This decision rule involves the calculation of one bandwidth only in contrast to two bandwidths required by (2.3). However, the rule (2.5) is not symmetric with respect to the ordering of the learning samples i.e. the order in which samples are considered influences the outcome of the rule. Moreover, the rule $\tilde{I}_{12}$ is based on the relationship between $\bar{f}_1$ and $\bar{f}_2$ which is satisfied only approximately by $\hat{f}_1$ and $\hat{f}_2$. Simulation experiments for parametric models discussed in Section 4 has shown that in most cases the mean probability of error for $\hat{I}$ lies inbetween the mean probability of error of $\tilde{I}_{12}$ and that of its analogue $\tilde{I}_{21}$ for which the samples $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are interchanged. That is why we focus on performance of the rule (2.3) in the paper.

3. An asymptotic representation

In Theorem 3.1 below, we prove an asymptotic expansion for $\hat{f}_1$. Calculation of the mean integrated squared error of its main part leads to a proposal of data-based bandwidths to be used in practical discriminant problems. Let $\pi_m := m/(m+n(m))$.

Theorem 3.1 *Assume that $\bar{f}_1$ satisfies Lipschitz condition in the neighbourhood of x such that $0 < x < 1$. Let K be compactly supported, twice differentiable kernel with a bounded second derivative on $\mathbb{R}^1$ and $|\pi_m - \pi| = o(m^{-1/2})$. Then the following representation holds*

$$\hat{f}_1(x) = \frac{1}{mb_m} \sum_{i=1}^{m} K\left(\frac{x - H(X_i)}{b_m}\right) - \pi \bar{f}_1(x) \left\{ \frac{1}{mb_m} \sum_{i=1}^{m} K\left(\frac{x - H(X_i)}{b_m}\right) - 1 \right\}$$

$$- (1-\pi) \bar{f}_1(x) \left\{ \frac{1}{nb_m} \sum_{i=1}^{n} K\left(\frac{x - H(Y_i)}{b_m}\right) - 1 \right\} + o_p\left((mb_m)^{-1/2}\right) \tag{3.1}$$

provided that b_m is a deterministic sequence such that $b_m \to 0$ and $mb_m^3 \to \infty$.

Theorem 3.1 is proved in the Appendix.

Remark 1 The smoothness assumptions for the kernel K are crucial for the basic expansion of $\hat{f}_1$ used in the proof to hold true. Thus the Theorem 3.1 in its present form does not cover the case of a uniform kernel.

Corollary 1 *Under assumptions of Theorem* 3.1

$$(mb_m)^{1/2}\left(\hat{f}_1(x) - b_m^{-1} \int K((x-y)/b_n) \, \mathrm{d}F_1 H^{-1}(y) \right)$$

$$\xrightarrow{\mathscr{L}} N\left(0,\ A\left(\pi,\ \bar{f}_1(x)\right) \int K^2(s) \, \mathrm{d}s\right),$$

where $A(\pi, x) = x(1 + (1-3\pi)x + \pi(2\pi - 1)x^2)$ for $x \in \mathbb{R}^1$.

Proof Note that the sum of the first and the second term in (3.1) is independent of the third term. Moreover, they are asymptotically normal with asymptotic variances

$$\left(1-\pi\bar{f}_1(x)\right)^2\bar{f}_1(x)\int K^2/mb_m \text{ and } (1-\pi)^2\bar{f}_1(x)^2\bar{f}_2(x)\int K^2/mb_m,$$

respectively (see e.g. Parzen (1962)). From this and the fact that $\pi\bar{f}_1(x)+(1-\pi)\bar{f}_2(x)=1$ Corollary directly follows.

Remark 2 *Note that for* $\pi=1/2$ $A(\pi,\ x)$ *simplifies to* $x(1-x/2)$.

Put $\tilde{h}(x)=\bar{f}_1H(x)$.

Corollary 2 *Under assumptions of Theorem* 3.1

$$(mb_m)^{1/2}\left(\hat{\bar{f}}_1(H_N(x))-b_m^{-1}\int K\left(\frac{H(x)-s}{b_n}\right)\,\mathrm{d}F_1H^{-1}(s)\right)$$

$$\xrightarrow{\mathscr{L}} N\left(0,\ A\left(\pi,\ \tilde{h}(x)\right)\int K^2(s)\ \mathrm{d}s\right),$$

provided that $0<H(x)<1$.

Proof The proof follows from the fact that an analogous representation to (3.1) holds for $\hat{\bar{f}}_1H_N(x)$; the only difference is that x on the RHS of (3.1) is replaced by $H(x)$.

Remark 3 Observe that

$$\tilde{h}(x)=f_1(x)/(\pi f_1(x)+(1-\pi)f_2(x)).$$

Thus, in the situation when F_2 is absolutely continuous w.r.t. F_1, an estimator of the density ratio $r(s)=f_2(s)/f_1(s)$ can be constructed as

$$\hat{h}(x)=T\left(\hat{\bar{f}}_1H_n(x)\right),$$

where $T(x)=(1-\pi)^{-1}(x^{-1}-\pi)$ for $x>0$. Using the delta rule and Corollary 2 it is easy to see that $\hat{h}(x)$ is asymptotically normal with asymptotic variance

$$(1-\pi)^{-2}\tilde{h}^{-4}(x)A\left(\pi,\ \tilde{h}(x)\right)\int K^2$$

$$=\left(r^2(x)+\pi r(x)+(1-\pi)r^3(x)\right)\int K^2/mb.$$

This result may be compared with the analogous property for the estimator of density ratio introduced in Ćwik and Mielniczuk (1989). They proved that the considered estimate is asymptotically normal with asymptotic variance $(r(x)+r^2(x))\int K^2/mb$. Thus the ratio of asymptotic variances is smaller or larger than 1 depending on whether $r(x)$ is smaller or larger than 1. It should be noted however that there is a difference in biases between the two estimates.

4. Choice of the smoothing parameter

We propose a method of choosing the bandwidth b for estimate $\hat{\bar{f}}_1$ of density $\bar{f}_1$ of the transformed random variable $H(X_1)$. The approach is based on the calculation of the mean integrated squared error for the approximation to $\hat{\bar{f}}_1$ given in (3.1). Thus the bandwidth choice corresponds to minimization of the global criterion for the performance of the density estimate, whereas our main concern is minimization of some misclassification probability measure (the mean probability of error in case of rule (2.1)). The last problem being intractable, we consider the proposed methodology as a first step in the discrimination context.

Assume that $\bar{f}_1$ is twice differentiable on $\mathbb{R}^1$ and K is symmetric. Then the mean integrated squared error for three terms on the RHS of (3.1) is equal to

$$R\left(\bar{f}_1''\right)\frac{b_m^4}{4}\left(\int z^2K(z)\,\mathrm{d}z\right)^2+\frac{1+(1-3\pi)R\left(\bar{f}_1\right)+\pi(2\pi-1)S\left(\bar{f}_1\right)}{mb_m}R(K)$$
$$+o\left(b_m^4\right)+o\left(\left(mb_m\right)^{-1}\right), \tag{4.1}$$

where $R(g)=\int g^2$ and $S(g)=\int g^3$ for arbitrary real function g. Assume from now on that $\pi=1/2$. Then, provided that $R(\bar{f}_1'')\neq 0$, a bandwidth b minimizing the two main terms in (4.1) is equal to

$$b_{\text{opt}}=\left(\frac{\left(1-R\left(\bar{f}_1\right)/2\right)R(K)}{R\left(\bar{f}_1''\right)\left(\int z^2K(z)dz\right)^2}\right)^{1/5}m^{-1/5}. \tag{4.2}$$

In numerical experiments reported here two estimators of b_{opt} are used as data-dependent bandwidths for the kernel estimate of $\bar{f}_1$. From now on K denotes the standard normal kernel.

Bandwidth b_{1c}.
This is a modified version of the bandwidth proposed by Silverman (1986) for the situation when an i.i.d. sample pertaining to $\bar{f}_1$ is available and is obtained assuming that $\bar{f}_1$ is Gaussian with mean a and variance σ^2. Then the following equalities hold

$$R\left(\bar{f}_1\right)=\left(2\sqrt{\pi}\,\sigma\right)^{-1},\qquad R\left(\bar{f}_1''\right)=3\left(8\sqrt{\pi}\,\sigma^5\right)^{-1},$$

and thus in the model

$$b_{\text{opt}}=\left\{\tfrac{4}{3}\left(1-\left(4\sqrt{\pi}\,\sigma\right)^{-1}\right)\right\}^{1/5}\sigma m^{-1/5}, \tag{4.3}$$

whereas the original bandwidth $\bar{b}_{\text{opt}}$ introduced by Silverman is equal to

$$(4/3)^{1/5}\sigma m^{-1/5}. \tag{4.4}$$

Data-based analogue b_1 of $\bar{b}_{\text{opt}}$ is obtained by plugging some estimator of σ in (4.4). The estimator $\hat{\sigma} = \min(s_m, \lambda/1.349)$, where s_m denotes the empirical standard deviation and λ is the interquartile range, is used. We define the empirical counterpart of b_{opt} by

$$b_{1c} = \begin{cases} \left\{\frac{4}{3}\left(1 - \left(4\sqrt{\pi}\hat{\sigma}\right)^{-1}\right)\right\}^{1/5} \hat{\sigma} m^{-1/5}, & \text{if } 4\pi\hat{\sigma} > 1; \\ b_1 & \text{otherwise.} \end{cases} \tag{4.5}$$

Bandwidth b_{2c}.

This is a modification of the Sheather-Jones method of bandwidth choice. The proposed method consists in seeking the solution to the equation (cf equation (2.2) in Sheather and Jones (1991))

$$\left\{ \frac{R(K)\left(1 - \hat{R}\left(\bar{f}_1\right)/2\right)}{\left\{\int z^2 K(z)dz\right\}^2 \hat{S}_D(\alpha_2(b))} \right\}^{1/5} m^{-1/5} - b = 0,$$

where $\hat{R}(\bar{f}_1)$ and $\hat{S}_D(\alpha_2(b))$ are certain kernel estimators of $R(\bar{f}_1)$ and $R(\bar{f}_1'')$, the latter using bandwidth $\alpha_2(b)$. A form of the function $\alpha_2(b)$ for complete observability case was used. For details we refer to Sheather and Jones (1991). This type of bandwidth is particularly useful in case of multimodal densities. The bandwidth for estimate $\hat{f}_2$ of density f_2 of the transformed random variable $H(Y_1)$ is chosen analogously.

5. Numerical experiments

The following six methods were considered:

a) DF;
b) RDF1;
c) RDF2;
d) RLDF;
e) RQDF;
f) LDF.

Method RDF1 (rank density function 1) is the rank method introduced in the paper when the corrected Silverman's bandwidth b_{1c} defined in (4.5) is used for both transformed samples. RDF2 is defined analogously with the bandwidth b_{1c} replaced by the corrected Sheather-Jones bandwidth b_{2c}. RLDF is a method based on Linear Discriminant Function (LDF) applied to the samples $H_N(X_1), \ldots, H_N(X_m)$ and $H_N(Y_1), \ldots, H_N(Y_n)$; RQDF is a rank analogue of

Quadratic Discriminant Function (cf e.g. Conover and Iman (1980)). Density function (DF) method is an empirical Bayes rule defined as follows

$$\hat{I}(Z)=\begin{cases}1, & \text{if } \pi\hat{f}_1(Z)>(1-\pi)\hat{f}_2(Z);\\ 2, & \text{otherwise,}\end{cases}$$

where $\hat{f}_1$ is the kernel estimate of the density function f_1 based on $X_1,\ldots,X_m$ and with the Silverman's bandwidth b_1 defined in Section 4. Estimate f_2 is defined analogously for the sample $Y_1,\ldots,Y_n$. The aim of considering DF and LDF together with the rank methods b) – e) is to measure the impact of information loss due to replacing original samples by their ranks in the considered discriminant problems.

Parametric models.

In all examples, F_1 is the distribution function of the standard normal distribution. The following distribution functions F_2 were considered:

I) $N(0.5, \sigma)$, for $\sigma=0.2, 0.3,\ldots,1.0$;

II) $N(\mu, 1)$, for $\mu=0.5, 0.6,\ldots,1.5$;

III) $C(p)$, for $p=0, 0.1,\ldots,1$, where $C(p)$ denotes Cauchy distribution with location parameter equal to p;

IV) a) $\frac{2}{3}N(0,1)+\frac{1}{3}N(0,(\frac{1}{10})^2)$;
b) $\frac{1}{2}N(-\frac{3}{2},(\frac{1}{2})^2)+\frac{1}{2}N(\frac{3}{2},(\frac{1}{2})^2)$;
c) $\frac{3}{4}N(0,1)+\frac{1}{4}N(\frac{3}{2},(\frac{1}{3})^2)$.

The corresponding densities for the examples in IV) are shown in Figure 1. They are called kurtotic unimodal, separated bimodal and skewed bimodal density in Marron and Wand (1992) (examples No. 4, No. 7, No. 8).

For each model 50 samples pertaining to F_1 and F_2 are generated for $m=n=25$ and 50, except for model III) in which case only $n=50$ is used. For

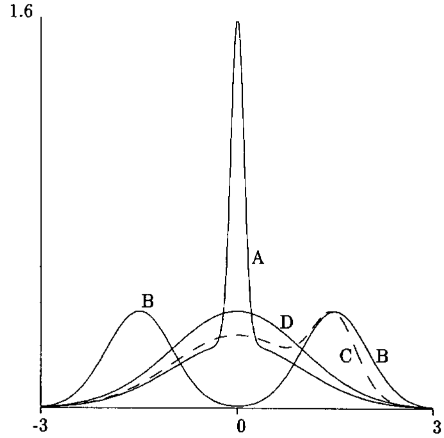

Fig. 1. Plot of the densities of the following distributions: A. $\frac{2}{3}N(0,1)+\frac{1}{3}N(0,0.1^2)$, B. $\frac{1}{2}N(-1.5,0.5^2)+\frac{1}{2}N(1.5,0.5^2)$, C. $\frac{3}{4}N(0,1)+\frac{1}{4}N(1.5),(\frac{1}{3})^2)$, D. $N(0,1)$.

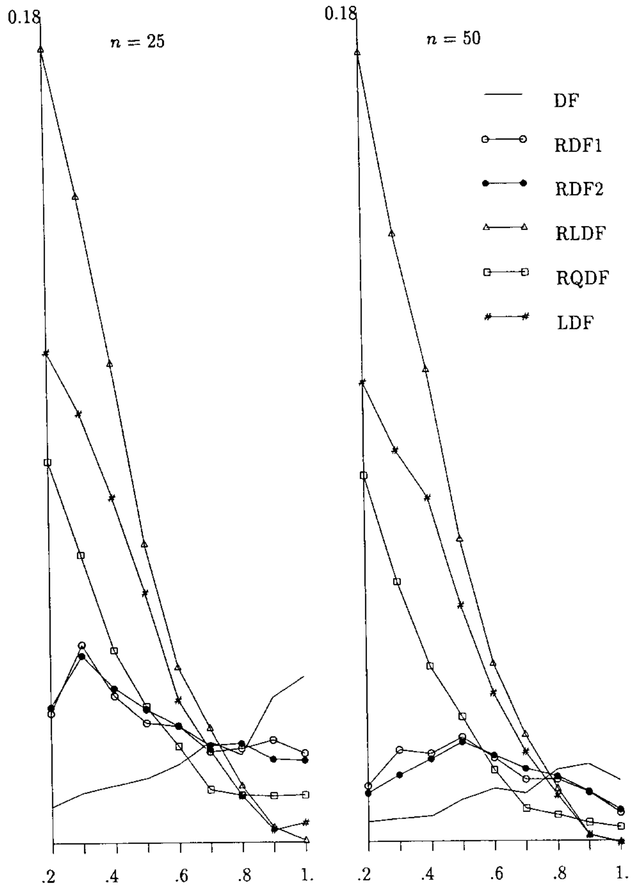

Fig. 2. Empirical MPE minus Bayes probability of error for six discriminant rules in the case of $N(0.5, \sigma)$, $\sigma = 0.2, 0.3, \ldots, 1.0$, sample size $n = 25, 50$.

each sample six discriminant rules a) – f) are constructed. Pertaining conditional probabilities of error are calculated using 10,000 elements generated from F_1 and F_2. A priori probability π is equal to 0.5 in all the examples. In Figures 2 – 6 and Tables the means and variances of the underlying histograms are displayed (in the figures, Bayes probability of error is subtracted from the mean probabilities of error). In the following paragraphs (I) to (IV) correspond to models (I) to (IV).

(I) For $n = 50$ and small σ, RQDF performs worse than RDF1 and RDF2; poor performance of RQDF in such a situation parallels that of QDF. LDF performs slightly better than RLDF but worse than RQDF. RDF1 and RDF2 perform similarly. DF method performs better than RDF1 for σ between 0.2 and 0.7 and worse than RDF1 for σ equal to 0.9 and 1.0. Results for $n = 25$ are similar to those for $n = 50$, when the mean probability of error (MPE) is concerned; variance is approximately two times larger than for $n = 50$.

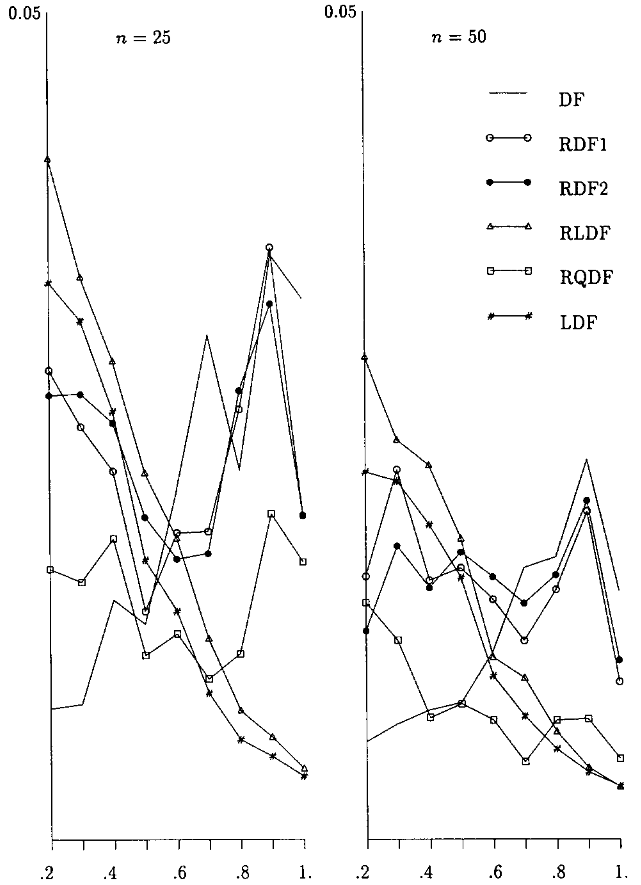

Fig. 3. Empirical variance of MPE for six discriminant rules in the case of $N(0.5, \sigma)$, $\sigma = 0.2, 0.3, \ldots, 1.0$, sample size $n = 25, 50$.

(II) Performance of RLDF is superior to all the remaining methods but LDF and very similar to that of LDF. This is understandable in view of discussion in Conover and Iman (1980) who indicated that RLDF is only slightly worse than LDF in this model. For μ ranging from 0.5 to 1 both RDFs perform better than DF.

(III) MPE for both RLDF and RQDF is larger than that of both RDFs. Performance of RLDF is disastrous for $p = 0$ and approaches that of RQDF for larger p. Observe however, that RLDF significantly improves over LDF. It is of interest to note that, in the case of $p = 0$ and $n = 50$, Hall and Wand (1988) obtained MPE = 0.3899 for empirical Bayes rule using specially designed cross-validation bandwidth, whereas the considered methods yield the following values of MPE: 0.4102 (DF), 0.4311 (RDF2), 0.4325 (RDF1), 0.4468 (RQDF), 0.4918 (LDF), 0.4990 (RLDF).

(IV) MPE for both RDFs is smaller than that of other methods in the case of models (a) and (b) and is slightly worse than that of DF. Of special interest is

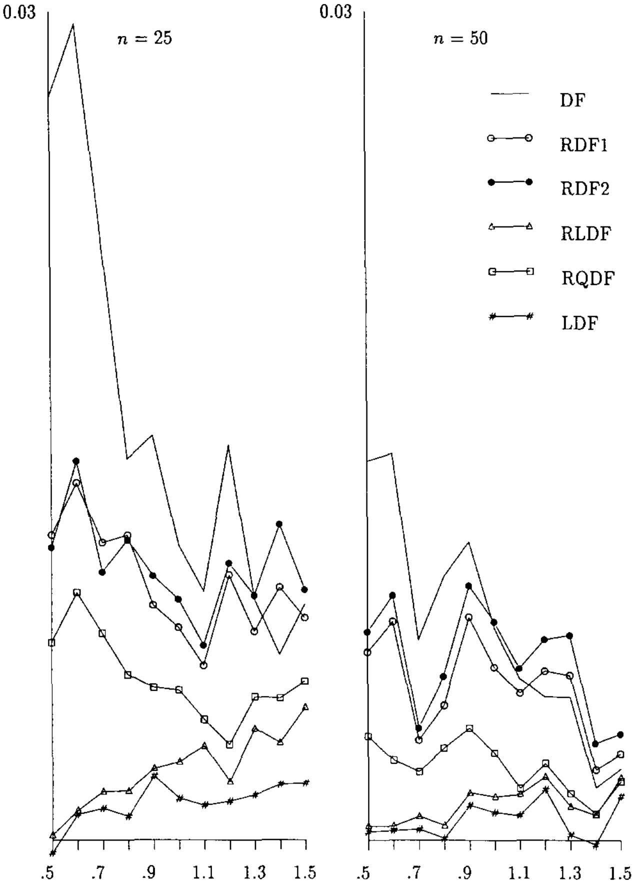

Fig. 4. Empirical MPE minus Bayes probability of error for six discriminant rules in the case of $N(\mu, 1)$, $\mu = 0.5, 0.6, \dots, 1.5$, sample size $n = 25, 50$.

situation (b) when both RDFs are better than DF. This is due to poor performance of Silverman's bandwidth for kernel estimate in case of densities with clearly separated modes.

In model (II) the Linear Discriminant Function is a direct empirical counterpart of the optimal Bayes rule (with only means estimated from the samples). The analogous statement is true for QDF in case of model (I). Rank counterparts RLDF and RQDF perform similarily and that is the main reason of superior performance of RLDF and RQDF to other rank methods in these models (apart from the situation for small σ in model (I). Rank density function approach produces slightly worse MPE for family (I) and (II) but exhibits much greater variability.

In the considered examples when F_2 is not normal, performance of RLDF and RQDF is not satisfactory and their MPE is substantially larger than that of rank density functions. On the other hand, general performance of DF is only

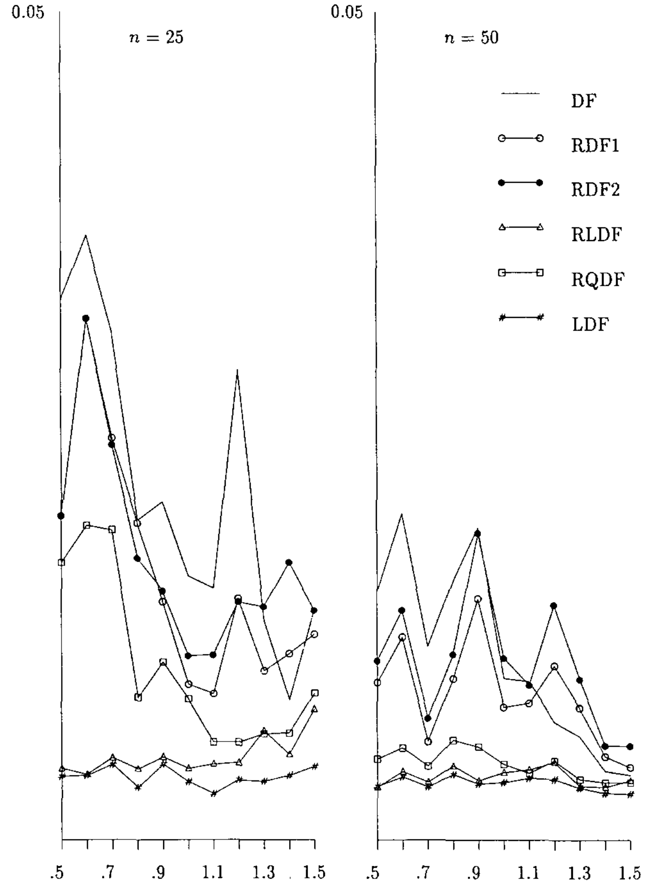

Fig. 5. Empirical variance of MPE for six discriminant rules in the case of $N(\mu, 1)$, $\mu = 0.5, 0.6, \ldots, 1.5$, sample size $n = 25, 50$.

slightly better than that of both RDFs; in some examples DF performs worse. The use of LDF and QDF is advocated when prior knowledge of approximate normality of underlying distributions is available. These rules should be replaced by RLDF and RQDF if we want to rely on rank data only. When significant departures from normality are expected, the rank discriminant function rule (2.3) provides a reasonable alternative to them among rank methods.

Appendix

Proof of Theorem 3.1. Observe that since $b_m \to 0$, K has a compact support and $0 < x < 1$, only the first summand in (2.4) is nonzero for sufficiently large m.

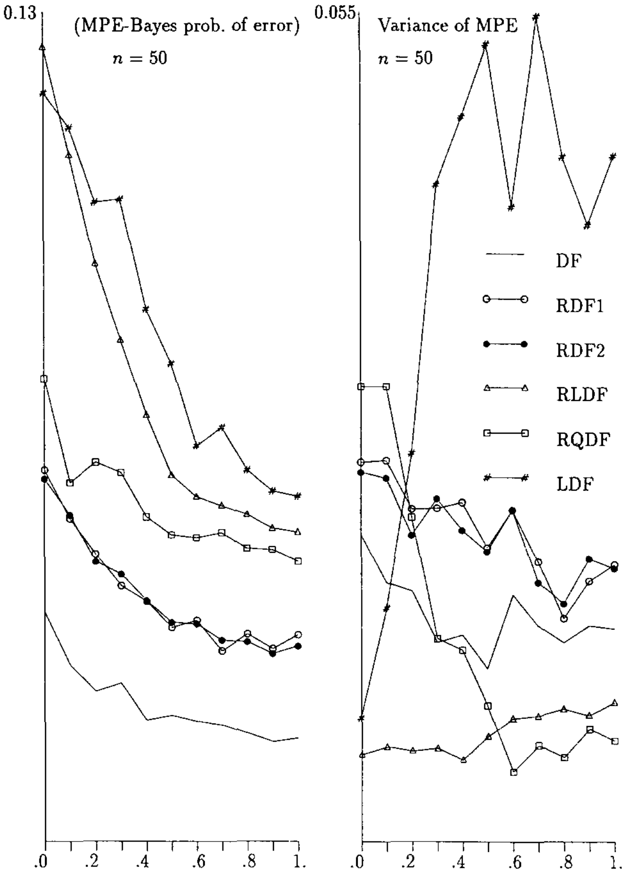

Fig. 6. Empirical MPE minus Bayes probability of error and empirical variance of MPE for six discriminant rules in the case of $C(p)$, $p = 0, .1, \ldots, 1$, sample size $n = 50$.

Consider the three terms Taylor decomposition of $\hat{f}_1$

$$\hat{f}_1(x) = \frac{1}{mb_m} \sum_{i=1}^{m} K\left(\frac{x - H(X_i)}{b_m}\right) - \frac{1}{mb_m^2} \sum_{i=1}^{m} K'\left(\frac{x - H(X_i)}{b_m}\right) \beta_N(X_i)$$

$$+ \frac{1}{2mb_m^3} \sum_{i=1}^{m} K''(\Delta_i) \beta_N^2(X_i) = I_0 + I_1 + I_2,$$

where $\beta_N(x) = H_N(x) - H(x)$ and Δ_i lies inbetween $(x - H_N(X_i))/b_m$ and $(x - H(X_i))/b_m$. It suffices to show that

$$I_2 = o_p\left((mb_m)^{-1/2}\right), \quad \text{and} \tag{A.1}$$

$$I_1 = -\frac{f_1(x)}{b_m} \int K\left(\frac{x - H(s)}{b_m}\right) d\beta_N(s) + o_p\left((mb_m)^{-1/2}\right). \tag{A.2}$$

Table 1
Empirical MPE for six discriminant rules in the case of models IV.a, IV.b and IV.c, sample size $n = 25, 50$

	DF	RDF1	RDF2	LDF	RLDF	RQDF
(a) $n = 25$	0.4247	0.4211	0.4204	0.4834	0.4929	0.4614
$n = 50$	0.4118	0.3960	0.3943	0.4913	0.4948	0.4575
(b) $n = 25$	0.2832	0.2580	0.2562	0.4890	0.4879	0.3773
$n = 50$	0.2577	0.2454	0.2458	0.4932	0.4795	0.3684
(c) $n = 25$	0.4244	0.4244	0.4238	0.4325	0.4284	0.4207
$n = 50$	0.4169	0.4173	0.4178	0.4269	0.4270	0.4164

The theorem follows from the above equalities after noting that I_1 differs from the sum of the second and the third term on the righthand side of (3.1) by $o(m^{-1/2})$ in view of the condition on π_m. In order to prove equality (A.1) observe first that a number of such i that $K''(\Delta_i) \neq 0$ is $O(mb_m)$ a.s.. To prove this statement note that $K''(\Delta_i) \neq 0$ implies $|x - H_N(X_i)| \leq Ab_m$ or $|x - H(X_i)| \leq Ab_m$, where $\operatorname{supp} K \subset [-A, A]$. Using Dvoretzky, Kiefer, Wolfowitz inequality and the condition $mb_m^2 \to \infty$ it is easy to see that the first condition implies $|x - H(X_i)| \leq (A+1)b_m$ a.s. Since the number of X_i satisfying this inequality is $O(mb_m)$ a.s., the assertion is proved. Using now boundedness of K'' and the fact that $\sup_x |\beta_N(x)| = O_p(m^{-1/2})$ we have that $|I_2| = O_p((mb_m^2)^{-1}) = o_p((mb_m)^{-1/2})$ provided that $mb_m^3 \to \infty$. We deal now with the term I_1. We prove that

$$E\left(I_1 - \tilde{I}_1\right)^2 = o\left((mb_m)^{-1}\right), \tag{A.3}$$

where $\tilde{I}_1 = -b_m^{-2} \int K'((x - H(s))/b_m) \beta_N(s) \, \mathrm{d}F_1(s)$. Observe that $(I_1 - \tilde{I}_1)^2$ is equal to

$$\begin{aligned} &\frac{1}{m^2 b_m^4} \sum_{i,j} K'\left(\frac{x - H(X_i)}{b_m}\right) K'\left(\frac{x - H(X_j)}{b_m}\right) \beta_N(X_i) \beta_N(X_j) \\ &- \frac{2}{mb_m^4} \sum_i K'\left(\frac{x - H(X_i)}{b_m}\right) \beta_N(X_i) \int K'\left(\frac{x - H(s)}{b_m}\right) \beta_N(s) \, \mathrm{d}F_1(s) \\ &+ \left\{ b_m^{-2} \int K'\left(\frac{x - H(s)}{b_m}\right) \beta_N(s) \, \mathrm{d}F_1(s) \right\}^2 . \end{aligned} \tag{A.4}$$

It is easy to see that the expectation of those summands in the first sum of (A.4) for which $i = j$ is $O((m^2 b_m^3)^{-1}) = o((mb_m)^{-1})$ provided $mb_m^2 \to \infty$. Moreover, the expectation of the remaining summands in the sum is equal to

$$\frac{m(m-1)}{m^2 b_m^4} \int K'\left(\frac{x - H(s)}{b_m}\right) K'\left(\frac{x - H(t)}{b_m}\right) w(s, t) \, \mathrm{d}F_1(s) \, \mathrm{d}F_1(t), \tag{A.5}$$

Table 2
Variances of MPE for six discriminant rules in the case of models IV.a, IV.b and IV.c, sample size $n = 25$, 50

	DF	RDF1	RDF2	LDF	RLDF	RQDF
(a) $n = 25$	0.0339	0.0372	0.0387	0.0438	0.0330	0.0278
$n = 50$	0.0183	0.0182	0.0180	0.0358	0.0259	0.0110
(b) $n = 25$	0.0384	0.0235	0.0229	0.0353	0.0603	0.0348
$n = 50$	0.0165	0.0077	0.0122	0.0241	0.0427	0.0238
(c) $n = 25$	0.0222	0.0195	0.0201	0.0292	0.0107	0.0198
$n = 50$	0.0144	0.0161	0.0158	0.0062	0.0091	0.0165

where $w(s, t) = \mathrm{E}(\beta_N(X_i)\beta_N(X_j) \mid X_i = s, \ X_j = t) = A_m(s, t)A_m(t, s)$ with $A_m(s, t) = \mathrm{E}((1 - \pi_m)F_{2n}(s) - (1 - \pi)F_2(s) + ((m-2)/N)F_{1m-2}(s) - \pi F_1(s) + N^{-1}(1 + I\{t \le s\}))$ and $F_{1m-2}(t)$ denotes the empirical distribution function based on $\{X_k\}$ with X_i, X_j omitted. Further, using the conditions on b_m and π_m it follows that (A.5) is equal to

$$\frac{m(m-1)}{m^2} S_H(b_m) + o(S_H(b_m)), \qquad \text{where}$$

$$S_H(b) = \frac{1}{b^4} \int B_{m,n}(s, t) K'\left(\frac{x - H(s)}{b}\right) K'\left(\frac{x - H(t)}{b}\right) \mathrm{d}F_1(t)\, \mathrm{d}F_1(s)$$

with $B_{m,n}(s, t) = (1 - \pi)^2(F_2(t) \wedge F_2(s) - F_2(t)F_2(s))/n + \pi^2(F_1(t) \wedge F_1(s) - F_1(t)F_1(s))/m$. Reasoning similarily for the second and the third term in (A.4) we obtain

$$E\left(I_1 - \bar{I}_1\right)^2 \le \left(\frac{m(m-1)}{m^2} - 2 + 1\right) S_H(b_m) + o\left((mb_m)^{-1}\right).$$

Since it is easy to see that $S_H(b_n) = O(m^{-1}b_m^{-2})$, (A.3) is proved. Note that

$$\bar{I}_1 = -b_n^{-2} \int K'\left(\frac{x - t}{b_m}\right) \bar{\beta}_N(t)\, \mathrm{d}F_1 H^{-1}(t),$$

where $\bar{\beta}_N(t) = \beta_N H^{-1}(t)$ and H^{-1} denotes the quantile function pertaining to distribution function H.
Further observe that using $|\bar{f}_1(t) - \bar{f}_1(x)| = O(b)$ for $|t - x| \le Ab$, $\sup_s |\bar{\beta}_n(s)| = O_p(n^{-1/2})$ and the fact that $b^{-1}\int K'((x - t)/b)\, \mathrm{d}t = O(1)$, it is easily shown that $\bar{I}_1$ is equal to

$$-\bar{f}_1(x) b_m^{-2} \int K'\left(\frac{x - s}{b_m}\right) \bar{\beta}_N(s)\, \mathrm{d}s + O_p(m^{-1/2}). \tag{A.6}$$

Integration by parts yields that the first term in (A.6) is equal to the first term in (A.2).

References

Behnen, K., Neuhaus, G. and Ruymgaart, F. (1983), Two sample rank estimators of optimal nonparametric score-functions and corresponding adaptive rank statistics, *Ann. Statist.*, **11**, 1175–1189.

Broffitt, J.D. (1982), Nonparametric classification, *Handbook of Statistics*, **2** (P.R. Krishnaiah and L.N. Kanal, Eds.), 139–168.

Ćwik, J. and Mielniczuk, J. (1989), Estimating density ratio with application to discriminant analysis, *Comm. Statist. – Theor. Meth.*, **18**, 3057–3069.

Conover, W.J. and Iman, R.L. (1980), The rank transformation as a method of discrimination with some examples, *Comm. Statist. – Theor. Meth.*, **9**, 465–487.

Devroye, L. and Györfi, L. (1985), *Nonparametric Density Estimation, The L_1 view.* New York: Wiley and Sons.

Diggle, P. and Marron, J.S. (1988), Equivalence of smoothing parameter selectors in density and intensity estimation, *J. Amer. Statist. Assoc.*, **83**, 793–800.

Gasser, T., Müller, H. and Mammitzsch, V. (1985), Kernels for nonparametric curve estimation, *J. Royal Statist. Soc. B*, **47**, 238–252.

Hall, P. and Wand, M.P. (1988), On nonparametric discrimination using density differences, *Biometrika*, **75**, 541–547.

Hand, D.J. (1982), *Kernel Discriminant Analysis*, Chichester: Research Studies Press.

Marron, S. and Wand, M. (1992), Exact Mean Integrated Squared Error, *Ann. Statist.*, **20**, 712–736.

Parzen, E. (1962), On the estimation of a probability function and the mode, *Ann. Math. Statist.*, **33**, 1065–1076.

Randles, R.H., Broffitt, J.D., Ramberg, J.R. and Hogg, R.V. (1978), Discriminant analysis based on ranks, *J. Amer. Statist. Assoc.*, **73**, 379–384.

Silverman, B. (1986), *Density Estimation for Statistics and Data Analysis*, London: Chapman and Hall.

Schuster, E.F. (1985), Incorporating support constraints into nonparametric estimates of densities, *Comm. Statist. – Theor. Meth.*, **14**, 1123–1138.

Sheather, S.J. and Jones, M.C. (1991), A reliable data-based bandwidth selection method for kernel density estimates, *J. Royal Statist. Soc. B*, **53**, 681–690.