IOSR Journal of Mathematics (IOSR-JM)
e-ISSN: 2278-5728, p-ISSN: 2319-765X. Volume 19, Issue 2 Ser. II (Mar. – Apr. 2023), PP 01-08
www.iosrjournals.org

A Maximum Rank Sum Statistic and its Application to Real Data Set

Gajraj Singh[*], Rahul Solanki[1], Piyush Bhardwaj[2]
[*]*Discipline of Statistics, School of Sciences, Indira Gandhi National Open University, Delhi, India*
[1]*Department of Operational Research, University of Delhi, Delhi, India*
[2]*Department of Economics, School of Social Sciences,Doon University, Dehradun, Uttrakhand, India*

Abstract-*This paper deals with a linear discriminant function based on the development of a nonparametric statistic. The statistic is defined as a maximum rank sum statistic which maximizes the sum of the ranks associated with the observations from one sample, based on the distances of all observations from two samples to a hyperplane. The optimal hyperplane which gives the maximum rank sum statistic is considered as a discriminant function for the two-population discriminant analysis. Mixed-integer and linear programming formulations are then derived for obtaining this type of nonparametric statistic. An efficient algorithm for computing this statistic is developed for the special case in which both samples are two-dimensional. Monte Carlo studies are conducted to evaluate the performance of the discriminant function derived from the maximum rank sum statistic in comparison with some statistical discriminant functions. The results show that the new discriminant function is competitive in various noncontaminated situations and performs better in some contaminated situations.*

Keywords: *-Derivation of the Maximum Rank Sum Statistic,Monte Carlo studies.*

Date of Submission: 07-03-2023 Date of Acceptance: 20-03-2023

I. Introduction

This paper deals with a formulations and methods of obtaining a nonparametric statistic and discusses an example usage of this statistic for constructing a two-population discriminant function. The statistic to be considered may be defined through the solution of the following problem. Let $x_i,\ i = 1,......,n_1,\ y_j$, $j = 1,......,n_2,$ be the m-dimensional samples from continuous populations I and II, respectively. The problem is to determine an m-dimensional column vector $c = (c_1,.......,c_m)', -\infty < c_k < \infty$, $k = 1,.....,m,$ and not all $c_k\,'s$ equal zero, such that the statistic

$$S = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \Psi[(y_j - x_i)c] \qquad (1)$$

is maximized, where ; $\Psi(t) = \begin{cases} 1, & \text{if } t \geq 0, \\ 0, & \text{if } t < 0 \end{cases}$

The problem is equivalent to finding an m-dimensional hyperplane with coefficient vector *c* such that after ranking in ascending order the distances for all *x*'s and *y*'s to the hyperplane, the sum of the ranks associated with the *y*'s is maximized.The equivalence of these two problems can be illustrated as follows. Let $d_i = x_i c,\ i = 1,.....,n_1,$ and $e_j = y_j c,\ j = 1,.....,n_2,$ be the distances of x_i and y_j to a hyperplane with coefficient vector *c*. Let the *x*'s and *y*'s be ranked based on the combined values of d_i and e_j, ranking in ascending order. When $(y_j - x_i)c = e_j - d_i \geq 0, y_j$ hasa higher or equal rank than x_i.The value of $\sum_{i=1}^{n_1} \Psi(y_j - x_i)c$ then gives the number of times y_j has a higher or equal rank than *x*'s , $j = 1,.....,n_2$. Also, *S* then represents the total number of times the ranks of the *y*'s are higher or equal to the ranks of the *x*'s. Let *R* be the sum of the ranks of $y_1,.....,y_{n_2}$ among $x_1,\ ...,x_{n_1}$ and $y_1,.....,y_{n_2}$. The relationship between *S* and *R* is given by $S = R - n_2(n_2 + 1)/2$. Therefore, the vector *c* which gives the maximum value of *S*, denoted as S^*, also gives the maximum value of *R*. The statistic S^* is thus defined as the maximum rank sum statistic. The motivation of deriving this nonparametric statistic is to use the corresponding hyperplane as a linear discriminant function for the two-population discriminant analysis problem. The problem can be defined as follows. Let $x_i, i = 1,.......n_1$ and $y_j, j = 1,......., n_2,$ be the m-dimensional training samples from populations I and II, respectively. A linear decision rule with coefficient vector *c* and a constant $\lambda$is to be derived to classify a new observation *z* into one of the two populations: if $L(z) = zc + \lambda \geq 0$, classify *z* into population I; otherwise, classify *z* into population II. The objective is to find a decision rule such that some measure of the probability of misclassification is minimized. Anderson [1], Goldstein and Dillon [3], and Lachenbruch [4]. The hyperplane obtained from the maximum rank sum statistic is to be considered as a discriminant function, it should be noted that the value of λ is undetermined from the statistic. Therefore, the hyperplane cannot be used directly as a regular discriminant function. But this problem can be solved if we assume that the hyperplane passes through the origin (i.e., $\lambda = 0$) and use a ranking procedure to assign a new observation *z* as follows.

Step 1. Compute $d_i = x_i c,\ i = 1,....., n_1$ and $e_j = y_j\, c,\ j = 1,....., n_2,$ where d_i and e_j are the distances of x_i and y_j to the hyperplane with coefficient vector *c* and passing through the origin. For a new observation *z*, compute $h = zc,$ the distance of *z* to the hyperplane.

Step 2. Let γ_x be the rank of *h* among *h* and $d_i, i = 1,....., n_1$, ranking in descending order. Let γ_y be the rank of *h* among *h*and $e_j,\ j = 1,....., n_2$, ranking in ascending order.

Step 3. Compute $p_x = \gamma_x / (n_1 + 1)$ and $p_y = \gamma_y / (n_2 + 1)$.The ranking procedure is then of the form:

(i) if $p_x > p_y$ assign *z* to population I;

(ii) if $p_x < p_y$ assign *z* to population II; and

(iii) if $p_x = p_y$ use a non-ranking procedure to assign *z*.

In the above ranking procedure $p_x(p_y)$ represents the probability of having an observation at least as extreme as *z* in the direction of the $y's\ (x's\ )$.Therefore this procedure has the objective of minimizing the total probability of misclassification. When $p_x = p_y$, any non-ranking procedure (e.g. Fisher's linear discriminant function [1]), can be used to assign *z*.

The rank discriminant function obtained from the maximum rank sum statistic also has the following properties.

(1) If the two training samples are linearly separable, this rank discriminant function will classify all the observations correctly. This is because the largest possible rank sum will be obtained, if there exist a hyperplane such that, relative to the hyperplane, every observation in one sample has a higher rank than each of the observations in the other sample. This is one of the desirable properties that every discriminant function should have.

(2) The relative positions among the observations contribute significantly to the determination of this discriminant function in addition to the magnitude of the data. This is due to the characteristics of both the maximum rank sum statistic and the rank procedure. The existence of outliers in the samples will have less effect on this discriminant function than on some parametric discriminant functions. This property is also useful if the samples are contaminated. This discriminant function should be more robust against contamination than those which do not have this property. In section 2 some mathematical formulations and their solution procedures are suggested for the determination of the maximum rank sum statistic. Monte Carlo studies are reported in section 3 for the performance of the rank discriminant function from this statistic and several other discriminant functions. A summary is presented in section

II. Derivation of the Maximum Rank Sum Statistic

The problem (1) discussed in section 1 may be formulated as a mixed-integer programming problem as follows:

$$\text{minimize}\, T = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} P_{ij}$$

$$\text{subject to} \quad (y_j - x_i)c + KP_{ij} \geq 0, \; i = 1, ..., n_1, \; j = 1, ..., n_2, \qquad (2)$$

$$\sum_{k=1}^{m} c_k^2 > 0,$$

where*K* is an arbitrary large positive constant, $c = \left(c_1, \; ..., c_m\right)'$ are continuous variables unrestricted in sign, and

$$P_{ij} = \begin{cases} 1, & \text{if } (y_j - x_i)c < 0, \\ 0, & \text{otherwise.} \end{cases}$$

In other words, $P_{i,j}$ equals 1 if y_i has a smaller rank than x_i relative to the hyperplane with coefficients *c*. Therefore the vector *c* which minimizes *T* will also give the maximum rank sum statistic *S**. To avoid the difficulty of assigning the value of *K* for a given set of samples and nonlinearity associated with the last constraint, the problem may be reformulated as follows:

$$\text{minimize}\, T = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} P_{ij}$$

$$\text{subject to} \qquad \sum_{k=1}^{m} (y_{jk} - x_{ik})c_k + KP_{ij} \geq 0, \qquad i = 1, ..., n_1, \; j = 1, ..., n_2,$$

$$-1 + 2D_k \leq c_k \leq 1 - 2E_k, \qquad k = 1, ..., m, \qquad (3)$$

$$\sum_{k=1}^{m} D_k + \sum_{k=1}^{m} E_k = 1$$

where P_{ij} and *c* are the same as defined in model (2), D_k and E_k are also zero-one variables to restrict one C_k equal to +1 or – 1, and the other *c*'s within (–1, +1); *k* has a lower bound:

$$K^* = m \max_{i,\, j,\, k} (| y_{jk} - x_{ik} |).$$

Note that formulation (3) is very similar to the mixed-integer program given in [5] except that there are more constraints and zero-one variables in (3). Formulation (3) can, therefore, be solved by either a branch-and-bound algorithm directly or with the use of Benders' decomposition [2]. But because of the large number of zero-one variables involved in (3), the use of those solution procedures is quite difficult except when the sample size is very small. An efficient algorithm for the two-characteristic problem (*m* = 2), however, can be developed as follows.

Let $s_{(1)} \leq ...... \leq s(N), N = n_1 n_2$, be the ordered slopes of the lines each of which passes through a combination of one *x* and one *y*. Assume that a rank sum statistic *S* is computed from a line with slope *s* such that $s_{(1)} < s < s_{(i+1)}, \; i \in (1, ..., N-1)$. If the position of the line is changed, the value of *S* will be the same as long as the slope of the line is still between $s_{(i)}$ and $s_{(i+1)}$.

Therefore, in this case there are at most (N + 1) situations which may give different values of the rank sum statistic. In addition, since the intercept of a line has no effect on the value of the statistic, without loss of generality we may assume that the line to be determined passes through the origin. The complete algorithm for the two-characteristic problem is now presented below.

Step 1. Initialize l = 0 and S* = 0. Calculate the slopes

$$s_t = (y_{j2} - x_{i2}) / (y_{j1} - x_{i1}),$$

$i = 1,...,n_1,\ j = 1,...,n_2$ where$t$ = 1,...,n_1n_2 = N denotes an index for each combination of i and j.

Step 2. Order the slopes $s_t,\ t = 1,....,N$, such that $s_{(1)} < ... < s(N)$..

Step 3. Let $ss_{(1)} = s_{(1)} - \delta,\ ss_{(t)} = \left(s_{(t)} + s_{(t+1)}\right)/2$), t = 2, ...,N, and $ss\left(N+1\right) = s_{(N)} + \delta,$ where δ is a small positive constant.

Step 4. Let $l = l + 1$. Compute the distances d_i and e_j of each x_i and y_j to the line with slope $ss_{(l)}$. The distances are proportional to the Eucledian distance from the points to the line.

Step 5. Rank the combined distances $d_1,........,d_{n_1}$ and $e_1,........,e_{n_2}$ in ascending order and compute the sum of the ranks of the y's, S, from the ranked distances. If S is greater than the current maximum rank sum S^*, let S^* = S. Go to step 4 if $l < N + 1$; otherwise, terminate.

To compute the distance d of a point $z = \left(z_1, z_2\right)$ to a line with slope s in step 4, the following simple formula may be used: $d = -z_1 s + z_2$

A software R program has been developed for the above algorithm. For simplicity only the first optimal slope is recorded if multiple solutions exist. Some limited computational experience indicates that this algorithm can solve problems with almost any sample size in a reasonable amount of computing time.

To obtain a statistic which has similar properties as the maximum rank sum statistic S* but is computationally feasible for problems with more than two characteristics (m > 2), a new formulation has to be developed. One of the alternatives which may provide a reasonable statistic is as follows. For any hyperplane with coefficient vector c, let $t_{ij} = (x_i - y_j)c$ be the difference of the distances from x_i and y_j to the hyperplane. Note that y_j will have a smaller rank than x_i if and only if $t_{ij} > 0$. To maximize the rank sum of the y's, a penalty t_{ij} may be assigned to the $x_i\,y_j$ pair if $t_{ij} > 0$. Therefore, a statistic may be developed by determining a hyperplane such that the total penalty is minimized. The problem may be formulated as follows:

minimize $T_1 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} t_{ij}$; subject to $\sum_{k=1}^{m}(x_{ik} - y_{jk})(c_k - t_{ij}) \le 0,\quad i = 1,...,n_1,\ j = 1,...,n_2$ (4)

$\sum_{k=1}^{m} c_k^2 = 1, t_{ij} \ge 0,\quad c_k$, unrestricted in sign.

The equality constraint in (4) is required to ensure that the trivial solution $c_k = 0,\ t_{ij} = 0$, for all i, j, and k, is infeasible. Note that although statistics T_1 and T are not the same, T_1 also has the property that the y's will tend to have larger ranks than the x's. To avoid the difficulty of solving formulation (4) due to the existence of a nonlinear constraint, a small-value "threshold" δ(<0) may be used to replace each zero on the right-hand side of the first (n_1 + n_2) constraints. In this way the trivial null solution will be infeasible and the formulation becomes a linear program.

The new formulation may be written as follows:

minimize $T_2 = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} t_{i\,j}$

subject to $\sum_{k=1}^{m}(x_{i\,j} - y_{j\,k})c_k - t_{i\,j} \le \delta, \qquad i = 1,....,n_1,\ j = 1,....,n_2,$ (5)

$t_{i\,j} \ge 0,\ \ c_k$ unrestricted in sign.

Another statistic which might provide a good approximation to T can be obtained from the following formulation:

maximize $T_3 = \sum_{k=1}^{m}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}(y_{j\,k} - x_{i\,k})c_k = \sum_{k=1}^{m}\alpha_k c_k = \alpha' c$

subject to

$\sum_{k=1}^{m} c_k^2 = 1$ (6)

whereα and *c* are *m*-dimensional column vectors.

Table 1

Values of Skewness and Kurtosis of the five distributions used in the Monte Carlo studies.

Distribution	Skewness	Kurtosis
1	0.0	1.8
2	0.0	3.0
3	0.0	9.0
4	0.5	8.8
5	1.6	8.8

In formulation (6) the statistic T_3 is obtained by maximizing the sum of differences of the distances from y_j and x_i to the optimal hyperplane with coefficients *c*. This problem may be solved quite easily by the method of Lagrange multipliers. The optimal solution of c_k is given by $c_k^* = \alpha_k / \|\alpha\|$. The difference between formulations (5) and (6) is that in (5) only the positive values of $(x_{i\,k} - y_{j\,k})c_k$ are considered in the objective function, whereas in (6) both positive and negative values are considered. To evaluate how well formulations (5) and (6) approximate formulation (3), a Monte Carlo study was conducted to compare the sum of the ranks associated with the *y*'s obtained from those formulations. Random samples of sizes 24 or 30 were generated for both *x*'s and *y*'s from five bivariate distributions in the lambda family (6,7). These distributions are characterized by their skewness and kurtosis and their values are given in table 1. The means for the *x* and *y* samples were at (0, 0) and (1, 1), respectively. Both samples had identity covariance matrices. For a given sample size and type of distribution, a pair of random samples for the *x*'s and *y*'s were generated and the rank sums of the *y*'s were computed from the solutions of formulations (3), (5), and (6). The process was repeated 100 times and the average rank sum from each formulation was computed. The ratios of the average rank sums among those three formulations for a sample size of 30 are shown in table 2. The result for a sample size of 24 is not given because it is very similar to the result given in table 2.

Table 2

Ratios of average rank sums from formulations (3),(5) and (6) for a sample size of 30.

Distribution	(3)/(5)	(3)/(6)	(5)/(6)
1	0.87	0.78	1.11
2	0.86	0.83	1.03
3	0.94	0.95	0.99
4	0.87	0.88	0.99
5	0.87	0.87	1.00

The result in table 2 shows that except for distribution 1 both formulations (5) and (6) approximate (3) quite well. In addition, the differences between (5) and (6) are negligible. Therefore, formulation (6) may be a better approximation to (3) since its optimal solution can be obtained more easily than formulation (5).

III. Monte Carlo studies

In this section the results of Monte Carlo studies for evaluating the performance of the rank discriminant function obtained from the maximum rank sum statistic, denoted as MRSS, in comparison with some statistical discriminant functions. The maximum rank sum statistic was computed using the algorithm for the bivariate data presented in section 2. The statistical discriminant functions included in the studies are the Fisher's linear discriminant function (FLDF) [1] and the LDF with Huber-type robust estimates of means and co-variances (LDF-Huber) [8]. The Monte Carlo studies include several noncontaminated and contaminated cases which are characterized by their respective bivariate distributions. In the noncontaminated cases the components of the bivariate random samples were generated from one of the five lambda distributions specified in table 1. In all those five cases the two populations had variances one and correlation zero. The mean of the first population, μ_1, was always at the origin, and the mean of the second population, μ_2, was at either (0.7071, 0.7071) or (1.0, 1.0) so that the Mahalanobis distance Δ^2 equals 1 and 2 respectively. In the contaminated cases, both the contaminating and contaminated distributions for each population were specified as normal and were assumed to have the same mean. The means for populations 1 and 2 were the same as the means in the noncontaminated cases. All the contaminated distributions had identity covariance matrices. The contaminating distributions had correlation zero and standard deviations 5, 10, or 20, representing mild, moderate, and severe contamination, respectively. The percentage of contamination was assumed to be 20% of the total sample size for both populations. The Monte Carlo studies proceeded as follows: First, training samples of sample size either 16 or 30 from each population were generated from a given set of populations. Various discriminant functions were derived from those training samples. Two new sets of samples, each of size 60, were then generated from the same populations and were classified by the various sample discriminant functions. The percentage of misclassification was recorded for each function. The process was repeated 100 times to compute the average percent of misclassification for each discriminant function. The whole process was then repeated with another set of populations. Since all the distributions in the contaminated cases were normal, the bivariate random variates were generated directly from the GGNQF routine in IMSL.

Table 3

Average % of misclassification for the non-contaminated cases with Δ^2 = 1 and sample size of 16.

Procedures	Distribution Type				
	1	2	3	4	5
FLDF	35.9	33.0	11.8	29.3	30.7
LDF-Huber	36.1	33.0	11.3	29.1	29.9
MRSS	33.5	32.8	13.3	29.6	30.2

Table 4

Average percentages of misclassification for the noncontaminated cases with Δ^2 = 1 and sample size of 30.

Procedures	Distribution Type				
	1	2	3	4	5
FLDF	33.1	31.3	12.2	28.4	29.7
LDF-Huber	33.3	31.4	11.4	28.1	29.4
MRSS	33.4	31.7	12.8	28.8	29.7

Table 5

Average percentages of misclassification for the noncontaminated cases with $\Delta^2 = 2$ and sample size of 16.

Procedures	**Distribution Type**				
	1	**2**	**3**	**4**	**5**
FLDF	28.6	25.2	7.0	21.1	22.9
LDF-Huber	28.9	25.0	6.6	21.8	22.0
MRSS	28.9	25.1	7.9	22.9	22.4

Table 6

Average percentages of misclassification for the noncontaminated cases with $\Delta^2 = 2$ and sample size of 30.

Procedures	**Distribution Type**				
	1	**2**	**3**	**4**	**5**
FLDF	26.5	24.5	6.9	21.3	22.4
LDF-Huber	26.6	24.7	6.5	20.8	21.6
MRSS	26.9	24.7	8.0	21.5	22.4

Table 7

Average percentages of misclassification for the noncontaminated cases with $\Delta^2 = 1$ and sample size of 16.

Procedures	**Contamination Level**		
	Mild	**Moderate**	**Severe**
FLDF	40.8	43.7	51.0
LDF-Huber	38.5	43.1	44.7
MRSS	40.9	40.5	40.5

Table 8

Average percentages of misclassification for the noncontaminated cases with $\Delta^2 = 1$ and sample size of 30.

Procedures	**Contamination Level**		
	Mild	**Moderate**	**Severe**
FLDF	38.6	45.0	49.4
LDF-Huber	37.4	39.1	41.2
MRSS	40.2	40.8	41.1

Table 9

Average percentages of misclassification for the noncontaminated cases with $\Delta^2 = 2$ and sample size of 16.

Procedures	**Contamination Level**		
	Mild	**Moderate**	**Severe**
FLDF	33.9	40.2	49.7
LDF-Huber	32.4	39.0	41.7
MRSS	35.9	36.5	36.0

Table 10

Average percentages of misclassification for the noncontaminated cases with $\Delta^2 = 2$ and sample size of 30.

Procedures	**Contamination Level**		
	Mild	**Moderate**	**Severe**
FLDF	33.5	39.7	46.7
LDF-Huber	33.3	35.2	36.7
MRSS	35.4	35.6	36.2

The average percentages of misclassification for the non-contaminated cases are given in tables 3-6. The standard errors of the averages are all less than 0.01. As shown in the tables, the three procedures performed equally well in all situations. The differences between them are not statistically significant in any of the cases.The results for the contaminated cases are given in tables 7-10. As expected, the LDF-Huber procedure is more robust against contamination than the regular LDF procedure. The RMRS procedure performed better than the LDF-Huber procedure for moderate and severe contamination with a sample size of 16, and the two procedures did equally well with a sample size of 30. On the other hand, the LDF-Huber procedure performed better than the RMRS procedure for mild contamination. The extent of overlapping between the two populations i.e. $\Delta^2 = 1$vs. $\Delta^2 = 2$ does not have a significant effect on the difference of the two procedures.

IV. Conclusions

A maximum rank sum statistic was defined and the methods of obtaining this nonparametric statistic were proposed. Some other nonparametric statistics, similar to this statistic but computationally more efficient, were also discussed and evaluated. This maximum rank sum statistic was then used to construct a linear discriminant function with the rank procedure. Results from a Monte Carlo study showed that the rank classification rule derived from this statistic was quite competitive for various noncontaminated situations and was robust against moderate and severe contamination.

If a quadratic or higher-order classification rule is to be derived from the maximum rank sum statistic, the solution will be considerably more difficult to obtain even for the two-variable case. It will be more practical if either the statistic T_1^* or T_2^* defined in section 2 is used to provide an approximate solution.

References

[1]. T.W. Anderson, Introduction to Multivariate Statistical Analysis (John Wiley & Sons, New York, 1958) ch. 6.

[2]. J.F. Benders, “Partitioning Procedures for Solving Mixed-Variables Programming Problems”, NumerischeMathematik 4 (1962) 238-252.

[3]. M. Goldstein and W.R. Dillion, Discrete Discriminant Analysis (John Wiley & Sons, New York, 1978).

[4]. P.A. Lachenbruch, Discriminant Analysis (Hafner Press, New York, 1975).

[5]. J.M. Liittschwager and C. Wang, “Integer Programming Solution of a Classification Problem", Management Science 24 (1978) 1515-1525.

[6]. J.S. Ramberg and B.W. Schmeiser, "An Approximate Method for Generating Symmetric Random Variables", Communications of the ACM 15 (1972) 987-990.

[7]. J.S. Ramberg and B.W. Schmeiser, "An Approximate Method for Generating Asymmetric Random Variables", Communications of the ACM 17 (1974) 78-82.

[8]. R.H. Randles, J.D. Broffitt, J.S. Ramberg and R.V. Hogg, "Generalized Linear and Quadratic Discriminant Functions Using Robust Estimates", Journal of the American Statistical Association 73 (1978) 564-568.

Gajraj Singh, et. al. "A Maximum Rank Sum Statistic and its Application to Real Data Set." *IOSR Journal of Mathematics (IOSR-JM)*, 19(2), (2023): pp. 01-08.