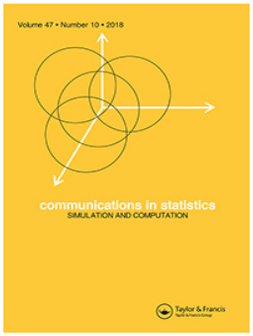

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Simulation and Computation

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lssp20

Robust quadratic discriminant analysis using S_n covariance

O. K. Sajana & T. A. Sajesh

To cite this article: O. K. Sajana & T. A. Sajesh (2021): Robust quadratic discriminant analysis using S_n covariance, Communications in Statistics - Simulation and Computation, DOI: 10.1080/03610918.2020.1868512

To link to this article: https://doi.org/10.1080/03610918.2020.1868512

Published online: 15 Mar 2021.

COMMUNICATIONS IN STATISTICS - SIMULATION AND COMPUTATION®
https://doi.org/10.1080/03610918.2020.1868512

Taylor & Francis
Taylor & Francis Group

Robust quadratic discriminant analysis using S_n covariance

O. K. Sajana and T. A. Sajesh

Department of Statistics, St Thomas' College (Autonomous), Thrissur, Kerala, India

ABSTRACT

This paper presents a robust method for robust estimation of quadratic discriminant analysis. The mean and covariance matrix for estimating quadratic discriminant rule is computed using a robust estimation method called S_n method established from a robust covariance estimator S_nCov. The performance of the proposed method is evaluated using the results of simulated samples. This outlier detection method is compared with some well-known methods available in the current literature. The application of the proposed method in real-life data is also executed in this paper.

ARTICLE HISTORY
Received 22 March 2020
Accepted 19 December 2020

KEYWORDS
Multivariate data; Quadratic discriminant rule; Robust estimation

1. Introduction

Discriminant Analysis (DA) is the multivariate technique that allows separating random objects into known groups of the population. The theory of discriminant function was introduced by Fisher (1938) for implementing the treatment of multiple measurements. The discriminant analysis can be considered as a statistical decision-making problem (Anderson 2004). The objective of discriminant analysis is the formulation of classification rules based on several training dataset and these determined rules are applied to classify the actual dataset. Discriminant analysis method includes Linear Discriminant analysis (LDA) and Quadratic Discriminant Analysis (QDA) for the assumptions according to equal and unequal population covariance matrices.

The classical methods of discriminant rules are often adopted to allocate multivariate observations to population groups and these are functions of sample mean vector and covariance matrix of the training dataset. Unfortunately, traditional rules are influential to outlying observations in the dataset which can mislead the classification of actual data. To overcome this situation, a robust alternative that is less sensitive to the presence of outlying observations are required for the estimation of parameters of discriminant rules.

Several multivariate robust estimation methods have been applied in literature for constructing robust quadratic discriminant rules. A robust estimation method called Minimum Covariance Determinant(MCD) proposed by Rousseeuw (1985) and Rousseeuw and Van Driessen (1999), that had been applied in robust quadratic and linear discriminant analysis for formulating discriminant rules by Hubert and Van Driessen (2004). They showed that the reweighting technique applied in MCD decreases the misclassification probabilities. Kurtosis method proposed by Peña and Prieto (2001) was implemented by Lakshmi and Sajesh (2018) in robust estimation of QDA parameters. Sajesh and Srinivasan (2019) developed robust QDA using comedian method and presented that the method is better than that of robust QDA using MCD and classical QDA. Various methods such as M-estimation (Maronna 1976), S-estimation (Lopuhaä 1989; Rocke 1996) and Orthogonalized Gnanadesikan-Kettenring (OGK) method (Maronna and Zamar 2002)

CONTACT O. K. Sajana sajana.kunjunni@gmail.com Department of Statistics, St Thomas' College (Autonomous), Thrissur, Kerala, India.

can also existed for robust estimation of location vector and scatter matrix. These methods can be adopted for the robust estimation of QDA parameters.

This article focuses on the study of Robust Quadratic Discriminant Analysis (RQDA) using the robust location and scatter based on S_n method discussed by Sajana and Sajesh (2020a). The effect of robust quadratic discriminant rules is investigated by comparing the overall misclassification estimate (MP) proposed by Hubert and Van Driessen (2004). The proposed robust QDA is compared with classical estimators and the RQDA's proposed by Hubert and Van Driessen (2004), Sajesh and Srinivasan (2019), Maronna (1976) and Maronna and Zamar (2002), to test the efficiency of the method. Moreover, real data applications are illustrated to ensure the performance of proposed RQDA in real life situations.

The generalization of classical QDA is discussed in the second section. The theoretical definitions are discussed in order to introduce the robust estimates of parameters of QDA. The third section consists of the definition of RQDA and misclassification probabilities. The results are tested for simulated training datasets as well as validation datasets and the estimated overall misclassification is described in section four. The next section contains the application of the proposed RQDA in real-life data. The results and findings are summarized in the last section.

2. Classical quadratic discriminant analysis

The theoretical generalization of classification procedure for discrimination with several groups of population $\pi_1, ..., \pi_k$ can be explained by considering the density $f_i(\mathbf{x})$ associated with population π_i to be multivariate normal with mean vectors $\boldsymbol{\mu}_i$ and covariance matrices Σ_i (Johnson and Wichern 1992). The derived discriminant rule for allocating the multivariate observation $\mathbf{x} \in \mathbf{R}^p$ to l^{th} population group is defined asallocate $\mathbf{x}$ to π_l if

$$\begin{aligned} \ln p_l f_l(\mathbf{x}) &= \ln p_l - \frac{p}{2}\ln(2\pi) - \frac{1}{2}\ln|\boldsymbol{\Sigma}_l| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_l)^T\boldsymbol{\Sigma}_l^{-1}(\mathbf{x}-\boldsymbol{\mu}_l) \\ &= \max_{i=1,...,k} \ln p_i f_i(\mathbf{x}) \end{aligned} \tag{1}$$

where p_i be the membership probability of population group π_i. The above discriminant rule can be simplified by ignoring the constant term, then the quadratic discriminant score will be

$$d_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\boldsymbol{\Sigma}_i| - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i) + \ln p_i \quad for\, i = 1, 2, ..., k \tag{2}$$

It is applied to find the discriminant rule with least total misclassification probability for normal population (Johnson and Wichern 1992). The discriminant rule is derived asallocate $\mathbf{x}$ to π_l if

$$d_l^Q(\mathbf{x}) = max\{d_1^Q(\mathbf{x}), d_2^Q(\mathbf{x}), ..., d_k^Q(\mathbf{x})\} \tag{3}$$

The quadratic discriminant score is reduced for the homogeneous population covariance matrices, it will be a linear combination of components of $\mathbf{x}$. Therefore the linear discriminant score is defined as

$$d_i^L(\mathbf{x}) = \boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\mathbf{x} - \frac{1}{2}\boldsymbol{\mu}_i^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}_i + \ln p_i \quad for\, i = 1, 2, ..., k \tag{4}$$

Practically, the scores are embodied of unknown parameters, $\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i$ and p_i (membership probabilities). Thus, sample mean vector $\bar{\mathbf{x}}_i$ and sample covariance matrix $\mathbf{S}_i$ of training datasets are adopted to compute the discriminant score explained in Equation (2). The estimated Classical Quadratic Discriminant Rule (CQDR) or QDA_C is then written asallocate $\mathbf{x}$ to π_l if

$$\hat{d}_l^{CQ}(\mathbf{x}) = max\{\hat{d}_1^Q(\mathbf{x}), \hat{d}_2^Q(\mathbf{x}), ..., \hat{d}_k^Q(\mathbf{x})\} \tag{5}$$

where $\hat{d}_i^Q(\mathbf{x})$ is defined as

$$\hat{d}_i^Q(\mathbf{x}) = -\frac{1}{2}\ln|\mathbf{S}_i| - \frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_i)^T\mathbf{S}_i^{-1}(\mathbf{x}-\bar{\mathbf{x}}_i) + \ln\hat{p}_i^C \quad for\ i = 1, 2, ..., k \tag{6}$$

The unknown membership probability can be estimated as a constant, *i.e.*, $\hat{p}_i^C = 1/k$ or can be estimated using relative frequencies of each population group, *i.e.*, $\hat{p}_i^C = n_i/n$ where $n = \sum_i^k n_i$. Since the classical discriminant function directly depends on the classical estimators of mean vector and covariance matrix of training data which are highly influenced by the presence of outliers, classification based on the classical discriminant function will be misleading. In order to solve the disparity in discrimination of observation due the presence of outliers, it is preferable to adopt robust estimators of mean vector and covariance matrix in the classification rule. The robust quadratic discriminant rule based on S_n estimators is described in the following section.

3. Robust quadratic discriminant analysis (RQDA)

Robust alternative for bivariate covariance (S_nCov) was proposed by Sajana and Sajesh (2020b) can be adopted to create initial estimates of robust covariance matrix for the iterative method of multivariate outlier detection and then estimation of location vector and scatter matrix.

The bivariate covariance estimate S_nCov of random variables X and Y is defined as,

$$S_nCov(X, Y) = \underset{i}{med}\{\underset{j\neq i}{med}\,[(x_i - x_j)(y_i - y_j)]\} \tag{7}$$

where $1 \leq i, j \leq n$ and *med* stands for high median ($[\frac{n}{2}]^{th}$ order statistic) for inner median, low median ($[\frac{n+1}{2}]^{th}$ order statistic) for outer median and n denotes number of samples. Sajana and Sajesh (2020), introduced a method for robust estimation of location vector and covariance matrix. For the purpose of robust estimation, robust covariance and correlation matrices by utilizing S_nCov is defined as,

$$\mathbf{COV}_{S_n}(\mathbf{X}) = (S_nCov(X_i, X_j)), \quad i, j = 1, 2, ..., p \tag{8}$$

Corresponding correlation matrix is defined as,

$$\boldsymbol{\xi}_{S_n}(\mathbf{X}) = \mathbf{DCOV}_{S_n}(\mathbf{X})\mathbf{D}^T \tag{9}$$

where $\mathbf{D}$ is diagonal matrix with diagonals $1/S_n(x_i), i = 1, ..., p$ and $S_n(X) = med_i\, med_j|x_i - x_j|$, for $i, j = 1, ..., n$, where *med* stands for low median ($[\frac{n+1}{2}]^{th}$ order statistic) for outer median and high median ($([\frac{n}{2}]+1)^{th}$ order statistic) for inner median. Since S_nCov lacks positive semi definiteness, an iterative technique introduced by Maronna and Zamar (2002) has been applied to make S_nCov positive definite and affine-equivariant. This iterative robust estimation procedure is termed as S_n *method*. These robust mean vector and covariance matrix estimates can be used to propose a RQDA.

The robust quadratic discriminant rule for RQDA is then defined as,allocate $\mathbf{x}$ to π_l if $\hat{d}_l^{RQ}(x) \geq \hat{d}_i^{RQ}(x)$ for all $i = 1, 2, ..., k$

$$\hat{d}_i^{RQ}(\mathbf{x}) = -\frac{1}{2}\ln|\hat{\boldsymbol{\Sigma}}_{i,S_n}| - \frac{1}{2}(\mathbf{x}-\hat{\boldsymbol{\mu}}_{i,S_n})^T\hat{\boldsymbol{\Sigma}}_{i,S_n}^{-1}(\mathbf{x}-\hat{\boldsymbol{\mu}}_{i,S_n}) + \ln\hat{p}_i^R \quad for\, i = 1, 2, ..., k \tag{10}$$

where $\hat{\boldsymbol{\mu}}_{i,S_n}$ and $\hat{\boldsymbol{\Sigma}}_{i,S_n}$ are the estimates of mean vector and covariance matrix using S_n method. The membership probability can be defined robustly by $\hat{p}_i^R = \tilde{n}_i/\tilde{n}$, where $\tilde{n} = \sum_{i=1}^n \tilde{n}_i$ and $\tilde{n}_i$ is the number of inliers in the i_{th} group. The performances of the RQDA based on S_n*method* (RQDA_{S_n}) proposed by Sajana and Sajesh (2020a) is then evaluated using estimated MP proposed by Hubert and Van Driessen (2004). The MP is defined as the weighted mean of the misclassification probabilities where weights are estimated membership probabilities:

$$MP = \sum_{i=1}^{k} \hat{p}_i^R MP_i \tag{11}$$

where MP_i be the misclassification probabilities. In this paper the evaluation of robust discriminant rules are conducted using R-programming language (R Core Team 2020). To ensure the performance of the proposed RQDA$_{S_n}$ is compared with the classical discriminant analysis and RDA based on MCD estimator, Comedian estimator, M-estimator and OGK estimator, using simulated samples.

4. Simulation results

The technique of MP includes splitting the observations randomly into two sets, one is the *training set* which is utilized for constructing discriminant rule and other set is the *validation set* which is used to estimate misclassification error. The estimated MP values for different case of contamination is discussed below.

The case A_p considers the uncontaminated data with dimension p where 500 observations from each population are drawn as training, which is denoted by

$$\begin{aligned} A_p.\pi_1 &: 500N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,p}) \\ \pi_2 &: 500N_p(\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{2,p}) \\ \pi_3 &: 500N_p(\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{3,p}) \end{aligned}$$

Training datasets which also contain outliers are samples from another distribution. These cases are given below.

$$\begin{aligned} B_p.\pi_1 &: 400N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,p}) + 100N_p(6\boldsymbol{\mu}_{1p}, \boldsymbol{\Sigma}_{4,p}) \\ \pi_2 &: 400N_p(\boldsymbol{\mu}_{2p}, \boldsymbol{\Sigma}_{2p}) + 100N_p(6\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{4,p}) \\ \pi_3 &: 400N_p(\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{3,p}) + 100N_p(6\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{4,p}) \end{aligned}$$

$$\begin{aligned} C_p.\pi_1 &: 800N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,p}) + 200N_p(6\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{4,p}) \\ \pi_2 &: 600N_p(\boldsymbol{\mu}_{2p}, \boldsymbol{\Sigma}_{2,p}) + 150N_p(6\boldsymbol{\mu}_{1p}, \boldsymbol{\Sigma}_{4,p}) \\ \pi_3 &: 400N_p(\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{3,p}) + \pi_3 : 100N_p(6\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{4,p}) \end{aligned}$$

$$\begin{aligned} D_p.\pi_1 &: 800N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,\mathbf{p}}) + 200N_p(6\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{4,\mathbf{p}}) \\ \pi_2 &: 400N_p(\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{2,\mathbf{p}}) + 100N_p(6\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{4,p}) \\ \pi_3 &: 400N_p(\boldsymbol{\mu}_{3p}, \boldsymbol{\Sigma}_{3,p}) + 100N_p(6\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{4,p}) \end{aligned}$$

$$\begin{aligned} E_p.\pi_1 &: 400N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,p}) + 100N_p(6\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{4,p}) \\ \pi_2 &: 450N_p(\boldsymbol{\mu}_{2p}, \boldsymbol{\Sigma}_{2,p}) + 50N_p(6\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{4,p}) \\ \pi_3 &: 350N_p(\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{3,p}) + 150N_p(6\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{4,p}) \end{aligned}$$

$$\begin{aligned} F_p.\pi_1 &: 160N_p(\boldsymbol{\mu}_{1,p}, \boldsymbol{\Sigma}_{1,p}) + 40N_p(6\boldsymbol{\mu}_{3,p}, 25\boldsymbol{\Sigma}_{4,p}) \\ \pi_2 &: 160N_p(\boldsymbol{\mu}_{2,p}, \boldsymbol{\Sigma}_{2,p}) + 40N_p(6\boldsymbol{\mu}_{1,p}, 25\boldsymbol{\Sigma}_{4,p}) \\ \pi_3 &: 160N_p(\boldsymbol{\mu}_{3,p}, \boldsymbol{\Sigma}_{3,p}) + 40N_p(6\boldsymbol{\mu}_{2,p}, 25\boldsymbol{\Sigma}_{4,p}) \end{aligned}$$

where $\boldsymbol{\mu}_{i,p}$ is the zero vector with i^{th} element equal to 1. The different choices of covariance matrix $\boldsymbol{\Sigma}$.

Table 1. Misclassification probability of $RQDA_{S_n}$, $RQDA_{Comedian}$ $RQDA_{MCD}$, QDA_C, $RQDA_M$ and $RQDA_{OGK}$ for $p = 10$.

	$RQDA_{S_n}$				$RQDA_{Comedian}$			
	MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
A_p	0.046	0.033	0.98	0.354	0.047	0.034	0.99	0.355
B_p	0.038	0.097	0.155	0.102	0.039	0.110	0.177	0.114
C_p	0.021	0.073	0.152	0.072	0.025	0.095	0.201	0.093
D_p	0.014	0.641	0.150	0.099	0.014	0.744	0.177	0.115
E_p	0.037	0.137	0.266	0.155	0.036	0.121	0.245	0.139
F_p	0.050	0.036	0.040	0.042	0.051	0.041	0.044	0.046
	$RQDA_{MCD}$				QDA_C			
	MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
A_p	0.046	0.034	0.9845	0.355	0.044	0.031	0.039	0.38
B_p	0.207	0.097	0.082	0.129	0.142	0.074	0.088	0.101
C_p	0.157	0.089	0.111	0.124	0.134	0.344	0.377	0.285
D_p	0.063	0.452	0.092	0.092	0.126	0.797	0.141	0.355
E_p	0.201	0.160	0.300	0.207	0.151	0.056	0.213	0.141
F_p	0.038	0.046	0.058	0.047	0.001	0.324	0.824	0.384
	$RQDA_M$				$RQDA_{OGK}$			
	MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
A_p	0.047	0.041	0.044	0.043	0.047	0.042	0.046	0.045
B_p	0.034	0.073	0.140	0.085	0.119	0.115	0.154	0.131
C_p	0.025	0.064	0.165	0.072	0.096	0.101	0.181	0.118
D_p	0.012	0.043	0.138	0.079	0.043	0.388	0.168	0.105
E_p	0.032	0.093	0.237	0.123	0.105	0.159	0.338	0.209
F_p	0.051	0.046	0.062	0.053	0.016	0.069	0.159	0.081

$$\begin{aligned}
\boldsymbol{\Sigma}_{1,3} &= diag(0.4, 0.4, 0.4)^2 \\
\boldsymbol{\Sigma}_{1,5} &= diag(0.4, 0.4, 0.4, 0.4, 0.4)^2 \\
\boldsymbol{\Sigma}_{1,10} &= diag(0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4, 0.4)^2 \\
\boldsymbol{\Sigma}_{1,20} &= diag(0.4, ..., 0.4)^2
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{2,3} &= diag(0.25, 0.75, 0.25)^2 \\
\boldsymbol{\Sigma}_{2,5} &= diag(0.25, 0.75, 0.25, 0.75, 0.25)^2 \\
\boldsymbol{\Sigma}_{2,10} &= diag(0.25, 0.75, 0.25, 0.75, 0.25, 0.75, 0.25, 0.75, 0.25, 0.75)^2 \\
\boldsymbol{\Sigma}_{2,20} &= diag(0.25, 0.75, 0.25, ..., 0.75)^2
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\Sigma}_{3,3} &= diag(0.9, 0.6, 0.3)^2 \\
\boldsymbol{\Sigma}_{3,5} &= diag(0.9, 0.6, 0.3, 0.9, 0.6)^2 \\
\boldsymbol{\Sigma}_{3,10} &= diag(0.9, 0.6, 0.3, 0.9, 0.6, 0.3, 0.9, 0.6, 0.3, 0.9)^2 \\
\boldsymbol{\Sigma}_{3,20} &= diag(0.9, 0.6, 0.3, ..., 0.6)^2
\end{aligned}$$

where *diag* stands for diagonal elements. Different situations of data contaminations are constructed using 20% outliers in B_p, C_p, D_p, E_p and F_p. From these various cases of training datasets, the case B_p contains equal number of observations and outliers. In the case of populations C and D, an unequal group size is considered. A varying outlier percentages are tested in trial dataset E and F. Each case is repeated for 100 Monte Carlo simulations and based on the inliers identified by the S_n method is then used to calculate relative frequencies of membership probabilities.

Tables 1 and 2 respectively shows average of total misclassification probabilities of 100 Monte Carlo experiments of each training groups and over all misclassification for $p = 10$ and $p = 20$.

Table 2. Misclassification probability of $RQDA_{S_n}$, $RQDA_{Comedian}$ $RQDA_{MCD}$, QDA_C, $RQDA_{OGK}$ and $RQDA_{OGK}$ for $p = 20$.

	$RQDA_{S_n}$				$RQDA_{Comedian}$			
	MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
A_p	0.021	0.012	0.051	0.016	0.024	0.116	0.99	0.344
B_p	0.082	0.0303	0.109	0.078	0.088	0.020	0.091	0.070
C_p	0.044	0.052	0.147	0.075	0.051	0.041	0.131	0.069
D_p	0.009	0.442	0.089	0.064	0.016	0.381	0.097	0.066
E_p	0.011	0.011	0.043	0.025	0.071	0.025	0.152	0.084
F_p	0.039	0.011	0.018	0.023	0.042	0.018	0.018	0.026
	$RQDA_{MCD}$				QDA_C			
	MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
A_p	0.024	0.013	0.996	0.344	0.022	0.009	0.014	0.015
B_p	0.276	0.052	0.027	0.121	0.094	0.016	0.025	0.045
C_p	0.196	0.092	0.039	0.131	0.083	0.321	0.348	0.251
D_p	0.084	0.782	0.032	0.110	0.064	0.842	0.085	0.330
E_p	0.302	0.187	0.313	0.250	0.101	0.012	0.152	0.088
F_p	0.036	0.031	0.032	0.033	0.012	0.481	0.754	0.411
	$RQDA_M$				$RQDA_{OGK}$			
	MP_1	MP_2	MP_3	MP	MP_1	MP_2	MP_3	MP
A_p	0.026	0.018	0.019	0.021	0.027	0.019	0.022	0.023
B_p	0.082	0.014	0.073	0.059	0.155	0.064	0.122	0.117
C_p	0.061	0.023	0.087	0.056	0.112	0.064	0.147	0.107
D_p	0.021	0.594	0.074	0.065	0.058	0.456	0.118	0.102
E_p	0.066	0.014	0.145	0.075	0.138	0.038	0.164	0.114
F_p	0.039	0.026	0.028	0.031	0.007	0.067	0.119	0.064

The results tabulated on both tables shows that the group wise misclassification probability and the over all misclassification of $RQDA_{S_n}$ is less in all cases compared to $RQDA_{MCD}$, QDA_C, $RQDA_M$ and $RQDA_{OGK}$. In comparison with $RQDA_{Comedian}$, $RQDA_{S_n}$ have less misclassification measurements in most of the cases. In Table 2 the misclassification decreasing rate increases for increase in the dimension as compared to $RQDA_{Comedian}$. In the case of comparison of the proposed method with $RQDA_M$, $RQDA_{S_n}$ shows better and rarely equal efficiency.

Apart from unequal mean vector and covariance matrix which is used in the previous simulation, unequal mean vector and equal covariance matrix and vice versa presented by Croux and Dehon (2008) are considered in the following simulation setup. For unequal mean vector and equal covariance matrix structure, two populations each consisting of 500 observations from $N_p(-1, I)$ and $N_p(1, I)$ are generated for *validation set* and additional 10% outliers from $N_p(9, I)$ and $N_p(-9, I)$ are generated for constructing *training* set. Similarly for equal mean vector and unequal covariance matrix structure, two populations each consisting of 500 observations from $N_p(0, 100I)$ and $N_p(0, I)$ are generated for *validation set* and additional 10% outliers from $N_p(0, I)$ and $N_p(0, 100I)$ are generated for constructing *training set*, where mean vector 1 stands for column vector with all elements equal to one and I denotes for identity matrix. The estimated misclassification probabilities of the two different group structures is displayed in Tables 3 and 4. The table values shows that the proposed RQDA performed better than the other compared methods with very low misclassification probabilities.

Moreover, Krzyśko and Smaga (2020), discussed different types of contaminations such as t_3-distribution contamination, scale contamination, one-direction shift location contamination and radial location contamination. The multivariate observations $\boldsymbol{x}_{ij}$ for population groups $k = 3$, are generated in the following way:

$$\boldsymbol{x}_{ij} = \boldsymbol{\Phi}\boldsymbol{\alpha}_{ij} + \boldsymbol{e}_{ij} \tag{12}$$

where $i = 1, 2, 3, j = 1, \ldots, n_i, \boldsymbol{\Phi}$ is the matrix of basis functions and $\boldsymbol{\alpha}_{ij}$ are $5p$-dimensional random

Table 3. Misclassification probability of $RQDA_{S_n}$, $RQDA_{Comedian}$ $RQDA_{MCD}$, $RQDA_M$ and $RQDA_{OGK}$ for unequal mean vector and equal covariance matrix structure.

	$p = 10$			$p = 20$		
	MP_1	MP_2	MP	MP_1	MP_2	MP
$RQDA_{S_n}$	0	0	0	0	0	0
$RQDA_{Comedian}$	0.002	0.003	0.002	0.001	0.001	0.001
$RQDA_{MCD}$	0.002	0.003	0.003	0.002	0.003	0.002
$RQDA_M$	0.003	0.005	0.004	0.0034	0.004	0.004
QDA_{OGK}	0.006	0.006	0.006	0.003	0.005	0.004

Table 4. Misclassification probability of $RQDA_{S_n}$, $RQDA_{Comedian}$ $RQDA_{MCD}$, $RQDA_M$ and $RQDA_{OGK}$ for equal mean vector and unequal covariance matrix structure.

	$p = 10$			$p = 20$		
	MP_1	MP_2	MP	MP_1	MP_2	MP
$RQDA_{S_n}$	0.001	0.001	0.001	0	0.001	0
$RQDA_{Comedian}$	0.041	0	0.021	0.039	0	0.021
$RQDA_{MCD}$	0.025	0	0.013	0.008	0.001	0.004
$RQDA_M$	0.031	0	0.016	0.029	0	0.015
QDA_{OGK}	0.023	0	0.012	0.023	0	0.012

vectors. Vector of measurement errors, $\boldsymbol{e_{ij}} = (e_{ij1}, ..., e_{ijp})$, where $e_{ijp} \sim N(0, 0.25r_{ijz})$ and r_{ijz} is the range of the z^{th} row of the matrix $\boldsymbol{\Phi\alpha}_{ij}$, for $z = 1, ..., p$.

The matrix of values of basis functions $\boldsymbol{\Phi}$ is defined as $\boldsymbol{\Phi} = diag(\phi_1^T, ..., \phi_p^T)$ is a block diagonal matrix of $\phi_k^T = (\phi_{k1}, ..., \phi_{kB_k}), k = 1, ..., p$. The values of basis functions, ϕ_k^T are generated using the function *fourier*() from package *fda* in the R-program (Ramsay et al. 2017) by specifying the two arguments, one is the vector of p arguments(sequence of values in [0,1]) and another argument is the number of basis functions in the Fourier basis ($B_k = 5$). The results will provide $p \times B_k$ matrix with p Fourier basis and $\boldsymbol{\Phi}$ is constructed by choosing rows as ϕ_k^T *for* $k = 1, ..., p$.

The random vectors $\boldsymbol{\alpha_{ij}}$ are generated according to $\alpha_{ij} \sim N_{5p}(\boldsymbol{\beta}_i, \boldsymbol{I}_{5p})$, where $\boldsymbol{\beta}_1 = 0_{5p}, \boldsymbol{\beta}_2 = (3, 0, ..., 0)^T, \boldsymbol{\beta}_3 = (0, 3, ..., 0)^T$ (Todorov and Pires 2007). In order to study the performance of the proposed $RQDA_{S_n}$, different kinds of contaminations are included in the above mentioned data set. The outlying points are simulated according to the following way:

- Multivariate t_3-distribution contamination: $\boldsymbol{\alpha}_{ij} = \boldsymbol{T}_{ij}/\sqrt{C_{ij}/3}$, where $\boldsymbol{T}_{ij} \sim N_{5p}(\boldsymbol{\beta}_i, \boldsymbol{I}_{5p})$ and $C_{ij} \sim \chi_3^2$
- Scale contamination: $\boldsymbol{\alpha}_{ij} \sim N_{5p}(\boldsymbol{\beta}_{;i}, 50\boldsymbol{I}_{5p})$
- One-direction shift location contamination: $\boldsymbol{\alpha}_{ij} \sim N_{5p}(\boldsymbol{\beta}_i + 10Q_{5p}1_{5p}, (0.25^2)\boldsymbol{I}_{5p}), Q_{5p} = \sqrt{\chi^2_{5p,0.999}/(5p)}$
- Radial location contamination: $\boldsymbol{\alpha}_{ij} \sim N_{5p}(\boldsymbol{\beta}_i + 10Q_{5p}\boldsymbol{m}_{ij}/||\boldsymbol{m}_{ij}||, (0.25^2)\boldsymbol{I}_{5p})$ and $\boldsymbol{m}_{ij}$ is one of the 2^{5p} random diagonals $(\pm 1, \ldots, \pm 1)$

The data for i^{th} population group is generated by substituting the simulated values of $\boldsymbol{\Phi}, \boldsymbol{\alpha}_{ij}$ and $\boldsymbol{e_{ij}}$ for $j = 1, ..., n_i$ in (12). The three different parameters of $\boldsymbol{\alpha}_{ij}$ are considered for generating data points for different population groups. For simulating specific percentages (25% and 40%) of outlying points in the data set, $\boldsymbol{x}_{ij}$ is generated in a similar manner with changes in the values of $\boldsymbol{\alpha}_{ij}$ and $\boldsymbol{e_{ij}}$ according to the type of outliers and replace the inlier(true data points) data with these outlying points.

The misclassification probabilities using simulated data when each group consisting of equal number of observations ($n_1 = n_2 = n_3 = 100$) that containing different type of outliers mentioned above are presented in Tables 5–8, respectively. This simulation process is repeated 100 times to produce the results more accurately and its averages are presented in the tables. Table 1 shows group MP values of the RQDA rules using different robust methods, in this table the $RQDA_{S_n}$

Table 5. Misclassification probability of $RQDA_{S_n}$, $RQDA_{Comedian}$ $RQDA_{MCD}$, $RQDA_M$ and $RQDA_{OGK}$ for t_3-distribution contamination when $p = 20$.

	25% of contamination			40% of contamination		
	MP_1	MP_2	MP_3	MP_1	MP_2	MP_3
$RQDA_{S_n}$	0.292	0.233	0.293	0.293	0.251	0.285
$RQDA_{Comedian}$	0.284	0.224	0.280	0.344	0.289	0.3431
$RQDA_{MCD}$	0.350	0.315	0.343	0.381	0.359	0.389
$RQDA_M$	0.338	0.318	0.354	0.393	0.355	0.397
QDA_{OGK}	0.300	0.257	0.311	0.381	0.358	0.389

Table 6. Misclassification probability of $RQDA_{S_n}$, $RQDA_{Comedian}$ $RQDA_{MCD}$, $RQDA_M$ and $RQDA_{OGK}$ for scale contamination when $p = 20$.

	25% of contamination			40% of contamination		
	MP_1	MP_2	MP_3	MP_1	MP_2	MP_3
$RQDA_{S_n}$	0.244	0.201	0.244	0.294	0.255	0.298
$RQDA_{Comedian}$	0.177	0.169	0.644	0.204	0.199	0.855
$RQDA_{MCD}$	0.291	0.266	0.292	0.292	0.251	0.297
$RQDA_M$	0.279	0.248	0.267	0.304	0.256	0.312
QDA_{OGK}	0.324	0.253	0.314	0.420	0.378	0.433

Table 7. Misclassification probability of $RQDA_{S_n}$, $RQDA_{Comedian}$ $RQDA_{MCD}$, $RQDA_M$ and $RQDA_{OGK}$ for one-direction shift location contamination when $p = 20$.

	25% of contamination			40% of contamination		
	MP_1	MP_2	MP_3	MP_1	MP_2	MP_3
$RQDA_{S_n}$	0.246	0.216	0.244	0.290	0.260	0.298
$RQDA_{Comedian}$	0.244	0.204	0.246	0.331	0.354	0.333
$RQDA_{MCD}$	0.446	0.421	0.441	0.518	0.493	0.525
$RQDA_M$	0.441	0.423	0.447	0.507	0.496	0.523
QDA_{OGK}	0.404	0.462	0.410	0.521	0.564	0.526

Table 8. Misclassification probability of $RQDA_{S_n}$, $RQDA_{Comedian}$ $RQDA_{MCD}$, $RQDA_M$ and $RQDA_{OGK}$ for radial location contamination when $p = 20$.

	25% of contamination			40% of contamination		
	MP_1	MP_2	MP_3	MP_1	MP_2	MP_3
$RQDA_{S_n}$	0.282	0.221	0.271	0.331	0.299	0.335
$RQDA_{Comedian}$	0.254	0.225	0.268	0.327	0.294	0.333
$RQDA_{MCD}$	0.452	0.417	0.435	0.514	0.500	0.516
$RQDA_M$	0.432	0.414	0.447	0.515	0.492	0.528
QDA_{OGK}	0.313	0.264	0.324	0.407	0.331	0.403

has less misclassification probabilities compared to $RQDA_{MCD}$, $RQDA_M$ and $RQDA_{OGK}$. It has less probability compared to $RQDA_{Comedian}$ for higher level of contamination. In the case of Table 6 (scale contamination), the proposed estimates produced MP values less than that of compared methods and more stable than $RQDA_{Comedian}$ for different ranges of contamination. In the case of one-direction shift contamination and radial location contamination, $RQDA_{S_n}$ performed equivalent to $RQDA_{Comedian}$ and outperformed $RQDA_{MCD}$, $RQDA_M$ and $RQDA_{OGK}$.

5. Real life example

First example of Hemophilia data is considered to evaluate the performance of $RQDA_{S_n}$ in real life data. The data openly available from https://www.rdocumentation.org/packages/rrcov/versions/1.5-2/topics/hemophilia. This data consists of measurements of two variables on 75 women which contains, 45 hemophilia A carriers and 30 normal women, where the first variable

measures log(AHF activity) and second variable measures log(AHF-like antigen). Johnson and Wichern (1992) studied and analyzed the dataset.

To determine discriminant rules, 60% of randomly selected data points are considered as a training dataset and the robust covariance matrix is calculated. The estimated membership probabilities using inliers in the selected training set are $p_1^R = 0.4$ and $p_2^R = 0.6$. The remaining 40% observations are considered as validation set to compute misclassification probabilities and MP. The estimated group misclassification of group I, group II and over all misclassification (in percentages) using RQDA_{S_n}, $\text{RQDA}_{Comedian}$, RQDA_{MCD}, RQDA_M and RQDA_{OGK} respectively are (8, 11–13, 15–17, 20, 23) and (8, 14, 23). From the misclassification estimation, it is clear that the proposed method has less error in classification and it is equivalent to that of CQDA since the data is uncontaminated.

6. Summary

Discriminant analysis is related to the discriminant score and classification rule associated with it. Since the classical discriminant scores are highly sensitive to the presence of outliers in the dataset, a more efficient RQDA is proposed in this article. The RQDA is constructed based on the robust estimation procedure developed on the basis of S_n method.

The evaluation of the performance of RQDA_{S_n} is conducted using Monte Carlo simulation study and it is compared with $\text{RQDA}_{Comedian}$, RQDA_{MCD}, CQDA (QDA_C), RQDA_M and RQDA_{OGK}. The simulation study consists of different percentages of contamination in generated population groups. The second set of simulation consists different choices location and scatter combinations and the MP values of proposed RQDA are compared with some well known methods which are mentioned above. The empirical results of comparison show that the proposed robust discriminant rule performed better than other compared methods. The proposed RQDA is also applied in real dataset as well to understand the efficacy of the proposed method. All these investigations supports the use of RQDA_{S_n} for discriminant analysis in high dimensional datasets.

Acknowledgements

The authors are thankful to the reviewer for their valuable comments and efforts towards improving our manuscript.

References

Anderson, T. W. 2004. *An introduction to multivariate statistical analysis*. New York: John Wiley & Sons, Inc.

Croux, C., and C. Dehon. 2008. Robust linear discriminant analysis using S-estimators. *Canadian Journal of Statistics* 29 (3):473–92. doi: 10.2307/3316042.

Fisher, R. A. 1938. The statistical utilization of multiple measurements. *Annals of Eugenics* 8 (4):376–86. doi: 10.1111/j.1469-1809.1938.tb02189.x.

Hubert, M., and K. Van Driessen. 2004. Fast and robust discriminant analysis. *Computational Statistics & Data Analysis* 45 (2):301–20. doi: 10.1016/S0167-9473(02)00299-2.

Johnson, R. A., and D. W. Wichern. 1992. *Applied multivariate analysis*. New Delhi: Prentice-Hall of India Private Limited.

Krzyśko, M., and Ł. Smaga. 2020. Robust multivariate functional discriminant coordinates. *Communications in Statistics - Simulation and Computation* 49 (3):717–33. doi: 10.1080/03610918.2019.1580731.

Lakshmi, R., and T. A. Sajesh. 2018. Robust quadratic discriminant analysis using Kurtosis method. *Journal of Computer and Mathematical Sciences* 9 (12):1907–14.

Lopuhaä, H. P. 1989. On the relation between S-estimators and M-estimators of multivariate location and covariance. *The Annals of Statistics* 17 (4):1662–83.

Maronna, R. A. 1976. Robust M-estimators of multivariate location and scatter. *The Annals of Statistics* 4 (1): 51–67. doi: 10.1214/aos/1176343347.

Maronna, R. A., and R. H. Zamar. 2002. Robust estimates of location and dispersion for high-dimensional datasets. *Technometrics* 44 (4):307–17. doi: 10.1198/004017002188618509.

Peña, D., and F. J. Prieto. 2001. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics* 43 (3):286–310. doi: 10.1198/004017001316975899.

Ramsay, J. O., H. Wickham, S. Graves, and G. Hooker. 2017. FDA – Functional data analysis. R Package Version 2.4.7.

R Core Team. 2020. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rousseeuw, P. J. 1985. Multivariate estimation with high breakdown point. In *Mathematical statistics and applications*, eds. W. Grossmann, G. Pflug, I. Vincze, W. Wertz, vol. B, 283–97. Dordrecht: Reidel Publishing.

Rousseeuw, P. J., and K. Van Driessen. 1999. A fast algorithm for the minimum covariance determinant estimator. *Technometrics* 41 (3):212–23. doi: 10.1080/00401706.1999.10485670.

Rocke, D. M. 1996. Robustness properties of S-estimators of multivariate location and shape in high dimension. *The Annals of Statistics* 24 (3):1327–45. doi: 10.1214/aos/1032526972.

Sajana, O. K., and T. A. Sajesh. 2020a. Multidimensional outlier detection and robust estimation using *Sn* Covariance. *Communications in Statistics-Simulation and Computation*. Advance online publication. doi: 10.1080/03610918.2020.1725820.

Sajana, O. K., and T. A. Sajesh. 2020b. *Sn* covariance. *Communications in Statistics - Theory and Methods* 49 (24): 6133–38. doi: 10.1080/03610926.2019.1628275.

Sajesh, T. A., and M. R. Srinivasan. 2019. Robust quadratic discriminant rule using Comedian. *Research and Review: Journal of Statistics* 8 (2):41–47.

Todorov, V., and A. M. Pires. 2007. Comparative performance of several robust linear discriminant analysis methods. *REVSTAT Statistical Journal* 5:63–83.