Gram–Schmidt Orthonormalization based Projection Depth

Muthukrishnan R[1], Vadivel M[2], Ramkumar N[3]
[1]Associate Professor, Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu, India
[2, 3]Research Scholar, Department of Statistics, Bharathiar University, Coimbatore, Tamil Nadu, India

Abstract—*Gram-Schmidt Orthonormalization (GSO) Euclidean vectors based depth function is proposed to compute projection depth. The performance of GSO algorithm has been studied with exact and approximate algorithms, used in the associated estimator namelyStahel-Donoho (S-D) location and scatter estimators, for bivariate data.The efficiency of GSO algorithm is checked out by computing average misclassification error in discriminant analysis under real and stimulatingenvironment. The study concludesthat GSO algorithm based projection depth estimators performs well when compared with exact and approximate algorithms.*

Keywords— *Gram-Schmidt process, Projection depth, Robust Discriminant analysis.*

I. INTRODUCTION

Data depth is a function which quantifies the centrality of a point in a given data cloud. It is closely related to central regions or trimmed regions. It plays an important role in many notable fields of statistics,namely; data exploration, ordering, asymptotic distributions and robust estimation.Manydepth procedures have been developedin the past few decades, namely,half space depth [13], simplicial depth [7], regression depth [11] and projection depth [8],[18],[14].

Data depth has been used to compute multivariate measures of location and dispersion. In recent years data depth,based on projections has been increasingly studied and is mostly used in multivariate statistics. The essence of depth function in multivariate analyses is to measure the degree of centrality of a point relative to a data set. An analysis of multivariate statistical data is done by consideringeach univariate projection of the data. The main idea is, a central point is located in a multivariate data cloud only if it is located centrally in each univariate projection of this data cloud. The depth of a point in a multivariate data cloud is defined asthe minimum of the depths arising from the univariate projection of the data.

This paper is organized as follows. Section 2 describes the projection depth and associated estimator. Section 3 presents an abstract of various computational algorithms such as exact, fixed and random. The computational aspects of the proposed algorithm, namely, Gram-Schmidt Orthonormalization procedure is also furnished in the section. Section 4 examines the performance of the proposed GSO algorithm and critically compares the various algorithms of projection depth procedure through some examples. Section 5demonstrates the superiority of the GSO algorithm by applying it in the discriminant analysis under real and simulation environment.

II. PROJECTION DEPTH AND ITS ASSOCIATED ESTIMATOR

[14]Established a projection-based depth function, which has the highest breakdown point among all the existing affine equivariant multivariate location estimators and associated medians. The projection depth is appeared very favorable to robust statistics when compared with the others depth notions.It is due to the reason that all the desirable properties of the general statistical depth function defined in [18], namely, affine invariance, maximality at center, monotonicity relative to deepest point, and vanishing at infinity are satisfied.Also, it can induce the favorable estimators, such as Stahel-Donohoestimator and depth weighted means for multivariate data [15], [16].[15]Introduced the concept of multidimensional trimming on projection depth. Exact computation of bivariate projection depth and Stahel-Donohoestimator,furthermore, with a proper choices of (μ,σ) are formulated and studied by [17].Furthercomputing issues of projection depth and its associated estimators has studied by [10]. The basic idea of computing, projection depth is summarized given below.

Let μ (.) and σ (.)be univariate location and scale measures, respectively. Then the outlyingness of a point $x \in R^P$ with respect to distribution functions F of X in as[8], [14].

$$O(x,F) = \sup_{\|u\|=1} |Q(u,x,F)| \qquad (1)$$

where $Q(x,F) = \left(u^T x - \mu(F_n)\right)/\sigma(F_u)$ and F_u is the distribution of $u^T x$. If $u^T x - \mu(F_u) = \sigma(F_u) = 0$, and $Q(u,x,F) = 0$, which denotes the projection of $x$onto the unit vectors u. Note that the most popular outlying function robust choice of μ and σ is the median (Med) and the median absolute deviation (MAD). Let F_u be the distribution furthermore, the projection depth and its associated estimator depend on the robust choice of (*Med*, *MAD*), $Q(x, u, X^n)$ in (1) with respect to the unit vectors u in (1).

$$O(x,X^n) = \sup_{\|u\|=1} Q(u,x,X^n) \qquad (2)$$

where, $Q(u,x,X^n) = \dfrac{|u^T x - Med(u^T X^n)|}{MAD(u^T X^n)}$

Where u^T denotes the projection of x onto the unit vector u and $u^T X^n = \{u^T X_1, u^T X_2, \ldots, u^T X_n\}$. The projection depth value of a given point x ε R^p with respect to F can be defined as $PD(X,F) = \dfrac{1}{[1+O(X,F)]}$

The famous Stahel-Donoho location estimator [2], [12] i.e. the Projection Weighted Mean (PWM) and Projection Weighted Scatter (PWS) is given by

$$PWM(F) = \frac{\int x W_1(PD(x,F))F(dx)}{\int W_1(PD(x,F))F(dx)} \qquad (3)$$

$$PWS(F) = \frac{\int (x-PWM(F))(x-PWM(F))^T w_2(PD(x,F))F(dx)}{\int w_2(PD(x,F))F(dx)}$$

(4)

Where *PWM*(*F*) and *PWS*(*F*) is the aforementioned Stahel-Donoho location and scatter estimators, $w_2(.)$ denotes the weight function on [0, 1] based on projection depth outlying function (μ(*F*),σ(*F*)) as respectively. Note that the projection depth and its associated estimators awell defined, certain monotony conditions are required as follows

$$\int w_i(PD(x,F))F(dx) > 0,$$

$$\int \|x\|^i w_i(PD(x,F))F(dx) < \infty, i = 1,2.$$

With a finite sample $X^n = \{X_1, X_2, \ldots, X_n\}$ from X and F_n be the corresponding empirical distribution of F based in X^n.

III. GSO PROCEDURE OF COMPUTING PROJECTION DEPTH

In high dimensions, approximate algorithms include fixed and random direction procedures [17], [3] gives low efficiency. To enhance the efficiency and reduce the computational complexity, Gram-Schmidt orthogonal Euclidean vector based projections have been introduced to compute depth.

The fixed direction procedure uses fixed m directions which cut the upperhalfplane equally, and chooses the direction which can maximize (2). While random direction procedure randomly picks some m directions and chooses the optimal direction for computing the projection depth. The detailed computational steps are given by [3].An exact algorithm of computation of bivariate projection depth and the Stahel-Donoho estimator has been studied by [17]. Further, the simplified version of the computational procedure is given by [9].

In mathematics, particularly linear algebra and numerical analysis, the Gram-Schmidt process is a method for orthogonalnormalizationof vectors in an inner product space, most commonly the Euclidean space R^n equipped with the standard inner product. The Gram-Schmidt process takes a finite, linearly independent set $S = \{v_1, \ldots, v_k\}$ for $k \le n$ and generates an orthogonal set $S' = \{v_1, \ldots, v_k\}$ that spans the same k-dimensional subspace of R^n as S. The basic idea of Gram-Schmidt process is as follows: Let $u_1 = v_1$, and

$$u_k = v_k - \sum_{j=1}^{k-1} proj_{u_j}(v_k) \qquad (5)$$

where, $proj_u(v) = \dfrac{\langle u,v\rangle}{\langle u,u\rangle} u,$

Here, $\langle u,v\rangle = u^T v$, denotes the inner product of the vectors u and v.Also, $e_k = \dfrac{u_k}{\|u_k\|}$. The sequence $u_1,\ldots,u_k$ is the orthogonal vectors, and the normalized vectors $e_1,\ldots,e_k$ form an orthonormal set. The computation of the sequence $u_{1,\ldots,}u_k$ is known as Gram-Schmidt orthogonalization, while the computation of the sequence $e_{1,\ldots,}e_k$ is known as Gram-Schmidt orthonormalization as the vectors are normalized.

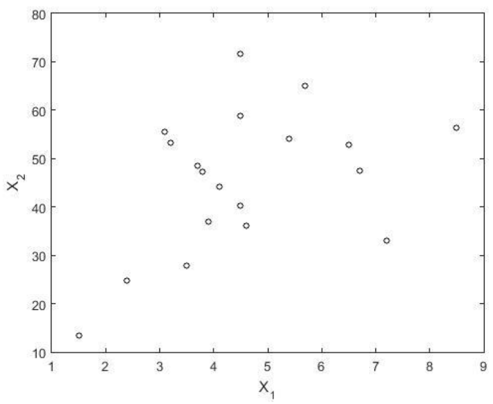
When this process is implemented on the vectors u_k are often not quite orthogonal, due to rounding errors.The Gram-Schmidt process can be stabilized by a small modification. Instead of computing the vector u_k as in (5), it is computed as $u_k^{(1)} = v_k - proj_{u_1}(v_k), u_k^{(2)} = u_k^{(1)} - proj_{u_1}(u_k^{(1)}), \ldots, u_k^{(k-1)} = u_k^{(k-2)} - proj_{u_{k-1}}(u_k^{(k-2)})$ each step finds a vector $u_k^{(i)}$ orthogonal to $u_k^{(i-1)}$ .Thus $u_k^{(i)}$ is also beingorthogonalized against any errors introduced in computation of $u_k^{(i-1)}$.

IV. NUMERICAL ANALYSIS

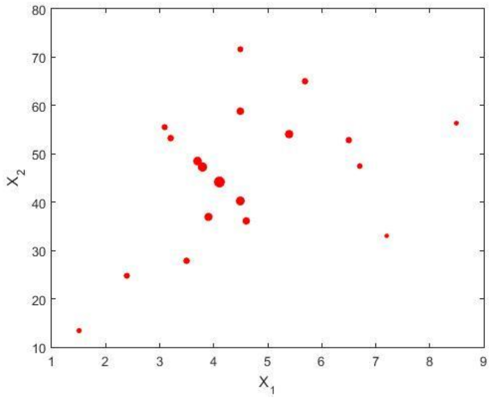
This section presents some examples to examine the performance of a Matlab implementation of the proposed GSO algorithm alongwith exact and approximate algorithms.

A. *Real Data*

To illustrate the performance of projection depth, a real data example is presented. The data set is taken from [5] (Sweat data, Page 215). The data consists of 19 observations. The data describe 19 healthy females were measured with two variables sweat rate (X_1) and sodium content (X_2). The scatter plot and the projection depth-size plots are displayed in the figure 1. It is noted that the larger size of the dot corresponds to the larger depth of the point. The computed projection depth values under the various algorithms with GSO are presented in the table 1.



(a)



(b)

Figure 1 (a) Scatter plot (b) Projection depth-size plot (GSO)

TABLE I
The Computed Projection Depth Values

Index	Fixed(1000)	Random(1000)	Exact	GSO
1	0.391675	0.386317	0.375349	0.491492
2	0.307042	0.306444	0.305747	0.306028
3	0.431026	0.426063	0.414516	0.555394
4	0.285616	0.279735	0.27095	0.311501
5	0.259398	0.253863	0.245614	0.278839
6	0.396943	0.396735	0.393545	0.401643
7	0.269014	0.262612	0.262133	0.2726
8	0.157949	0.160002	0.15457	0.1567
9	0.240117	0.243466	0.234637	0.237457
10	0.443961	0.415142	0.41369	0.467244
11	0.449288	0.449321	0.449288	0.457557
12	0.310127	0.312281	0.30121	0.415504
13	0.303121	0.303158	0.303121	0.314594
14	0.513441	0.513905	0.508065	0.518881
15	0.197541	0.192297	0.191923	0.198288
16	0.164966	0.167696	0.164595	0.173674
17	0.205882	0.208916	0.201178	0.266995
18	0.280616	0.282938	0.276772	0.296182
19	**0.582278**	**0.580748**	**0.568047**	**0.695879**

It is noted that, all procedures represent the 19th observation as the location of a given data, since it has the largest depth value. Further comparing the depth value under various algorithms, the GSO gives the highest among them.

B. Simulation Results

A simulation study is performed to compare the efficiency of the proposed GSO procedure along with various notions of projection depth procedure. The data (n=25) are generated from a multivariate normal distribution, mean vector μ= (0,0) and the unit covariance matrix, $\Sigma=I_2$. The results are listed in table 2.

TABLE II
The Computed Projection Depth Values

S.No	Fixed(1000)	Random(1000)	Exact	GSO
1	0.279751	0.279985	0.279751	0.284688
2	0.177699	0.177773	0.177699	0.187825
3	0.300351	0.300372	0.300242	0.312825
4	0.418745	0.419082	0.418745	0.421392
5	0.402708	0.402620	0.402544	0.406353
6	0.260813	0.261744	0.260813	0.262646
7	0.177237	0.177604	0.177137	0.187285
8	0.314549	0.315592	0.314549	0.316594
9	0.327815	0.327222	0.326425	0.340121
10	0.443497	0.443604	0.443497	0.46261
11	0.395648	0.395218	0.394840	0.402085
12	0.180137	0.180194	0.180076	0.193111
13	0.490847	0.491594	0.490847	0.493041
14	0.236695	0.236771	0.236622	0.25192
15	0.287932	0.288072	0.287932	0.297997
16	0.261341	0.262160	0.261341	0.262905
17	**0.557417**	**0.558160**	**0.557417**	**0.559598**
18	0.201091	0.201263	0.201091	0.201418
19	0.355026	0.355581	0.355026	0.366898
20	0.361799	0.361851	0.361799	0.369937
21	0.446930	0.446945	0.446930	0.475139
22	0.490172	0.490325	0.490172	0.510681
23	0.291422	0.291432	0.291422	0.309773
24	0.402583	0.402858	0.402583	0.404746
25	0.313248	0.313340	0.313248	0.327495

The table indicates that, 17th observation represents the location of generating data, since it has the largest depth value. Further, it is noted that GSO gives the highest depth value compared to the exact and approximate algorithms.

V. APPLICATION IN DISCRIMINANT ANALYSIS

The superiority of the GSO algorithm over the approximate and exact algorithms is studied by performing classification technique, namely discriminant analysis and compared the misclassification rate. [4] Developed the computational algorithm for fast and robust discriminant analysis, which is used for MATLAB implementation. The Stahel-Donoho estimator based projection depth approach is used for computing location and scatter values. Further, the analysis was performed simulated data with contamination.

A. Real Data

A real data set is taken from [6] (Page 584). The data consists of two different groups: π_1 is ridingmower owners and π_2 is without ridingmowers(non-owners) with each of sample 12. The owners or non-owners on the basis ofvariables, income (x_1) and lot size (x_2). The computed group-wise misclassification and its averages are presented in Table 3.

TABLE III
Computed Misclassification Probabilities
under Various Projection Depths

Procedures	Misclassification Probabilities		
	π_1	π_2	Average
Exact	0.1667	0.2500	0.2083
Fixed	0.1667	0.2500	0.2083
Random	0.1667	0.2500	0.2083
GSO	0.0833	0.1667	0.1250

It is observed that, the GSO algorithm gives very less average misclassification rate when compared with approximate and exact algorithms. That GSO procedure is misclassifies only 12%, but all other procedures misclassifies around 21% of the original data.

B. Simulation Result

To compare the GSO procedure with the approximate and exact procedure a simulation study is also performed with/without contamination. The data were generated from two different normal distributions (g=2, p=2) with varying sample sizes 50 and 100.The data were generated from thenormal distribution with covariance matrices $\sum_1$=$I_2$and $\sum_2$= 1.5$I_2$and means μ_1= (1, 1) and μ_2= (3, 3). The location and scale contaminations are applied as described using the values of μ_1= (-4, -4) and μ_2= (-5, -5) along with the covariance matrices $\sum_1$ =3 $I_2$and $\sum_2$ = 2I_2.The various levels of contamination such as 0%,5%, 10%, 15% and 20% were considered in two cases also. The obtained results with the contamination are displayed in the table4.

TABLE IV
Computed Misclassification Probabilities
under Various Contamination Levels

n_1=n_2=50				
Error	**Exact**	**Fixed**	**Random**	**GSO**
0.00	0.0645	0.0645	0.0645	0.0645
0.05	0.0690	0.0690	0.0690	0.0690
0.10	0.0769	0.0769	0.0769	0.0718
0.15	0.0854	0.0854	0.0854	0.0769
0.20	0.0897	0.0897	0.0897	0.0824
n_1=n_2=100				
0.00	0.0952	0.0952	0.0952	0.0952
0.05	0.0984	0.0984	0.0984	0.0984
0.10	0.1205	0.1205	0.1205	0.1193
0.15	0.1250	0.1250	0.1250	0.1236
0.20	0.1582	0.1582	0.1582	0.1429

It is noted that, when the contamination level increases misclassification probabilities is also increasing under all the procedures. On comparing the average probability of misclassification values in the above table, it is evident that the procedures GSO algorithm produces less when compared with exact and approximate algorithms. It is concluded that the GSO performs better than the exact and approximate algorithms. It shows that it is superior to the other algorithms under with/without contaminating data.

VI. CONCLUSION

This paper presents a novel idea of computing, projection depth. The computational aspect of the GSO algorithm is described. The performance of the GSO algorithm is discussed through numerical analysis. Further, the superiority of the GSO is demonstrated over the exact and approximate procedures by applying it in discriminant analysis under with/without contamination. It is concluded that, the performance of GSOprocedure is much better than approximate and exact algorithms. The study can be extended to higher dimensions. The new GSO procedure can be applied in almost all multivariate analysis and in turn it is very useful to research communities doing research in the field of data mining and computer vision.

ACKNOWLEDGEMENT

The authors would like to thank the anonymous reviewers and the Editor for their valuable comments on the paper, which helped us to improve the quality of the work.The research was partially supported by University Grant Commission, India under Rajiv Gandhi National Fellowship (RGNF) program and it was carried out atBharathiar University, Tamil Nadu and India.

REFERENCES

[1] Ben Noble and James Daniel.: *Applied Linear Algebra*, 3^{rd} *edition*. Prentice Hall, Englewood Cliffs, N.J., (1988).

[2] Donoho, D. L., and Gasko, M.: *Breakdown properties of location estimates based on halfspace depth and projected outlyingness*. The Annals of Statistics. 20, 1803–1827, (1992).

[3] Dyckerhoff, R,.:*Data depthssatistiying the projection property*. All emeinestatistischesArchiv, 88, 163-190, (2004).

[4] Hubert, M. and Van Driessen, K.: Fast and Robust Discriminant Analysis. Computaional Statistics and Data Analysis, 45,301,320, (2004).

[5] Johnson, R.A., and Wichern, D.W.: *Applied multivariate analysis,3rd ed.* Prentice Hall, Englewood Cliffs, New Jersey, (2007).

[6] Johnson, R.A., and Wichern, D.W.: *Applied multivariate analysis, 9^{th} ed.* Prentice Hall, Englewood Cliffs, New Jersey, (2009).

[7] Liu, R.Y.: On a notion of data depth based on random simplices. *The Annals of Statistics*, 18, 191-219, (1990).

[8] Liu, R.Y.: Data depth and multivariate rank test. In: L1-Statistical Analysis and Related Methods, 279-294. North-Holland, Amsterdam, (1992).

[9] Liu, X.H., Zuo, Y.J., Wang, Z.Z.: Exactly computing bivariate projection depth contours and median. Preprint, (2011).

[10] Liu, X., and Zuo, Y.: Computing projection depth and its associated estimators. Stat.Comput., 24, 51-63, (2014).

[11] Rousseeuw, P.J., and Hubert, M.: Regression depth. Journal of the American Statistical Association, 94, 388-433, (1999).

[12] Stahel, W.A.: Breakdown of covariance estimators. Research Report 31, 1029-1036, (1981).

[13] Tukey, J.W.: Mathematics and the picturing of data. In: Proceedings of the International Congress of Mathematicians, pp. 523-531. Canadian Mathematical Congress, Montreal, (1975).

[14] Zuo, Y. Projection-based depth functions and associated medians. The Annals of Statistics, 31, 1460-1490, (2003).

[15] Zuo, Y.J.: Multidimensional trimming based on projection depth. The Annals of Statistics, 34, 2211-2251, (2006).

[16] Zuo, Y.J., Cui, H.J., He, X.M.: On the Stahel-Donoho estimators and depth – weighted means for multivariate data. The Annals of Statistics, 32, 189-218, (2004).

[17] Zuo, Y.J., and Lai, S.Y.: Exact computation of bivariate projection depth and Stahel-Donoho estimator. Computational Statistics & Data Analysis, 55, 1173-1179, (2011).

[18] Zuo, Y., and Serfling, R.: General notions of statistical depth function. The Annals of Statistics, 28, 461–482, (2000).