Journal of Statistical Theory and Practice (2019) 13:41
https://doi.org/10.1007/s42519-019-0041-z

ORIGINAL ARTICLE

Nonparametric Control Charts Based on Data Depth for Location Parameter

M. S. Barale[1] · D. T. Shirke[1]

Abstract
Traditional control charts like Hotelling T^2 are based on the assumption of multivariate normality and also inapplicable to high-dimensional data. A notion of data depth has been used to measure centrality of a given point in a given data cloud. The data depth inferences do not require multivariate normality and any constraint on the dimension of the data. Liu (J Am Stat Assoc 90(432):1380–1387, 1995) provided control charts for a multivariate processes based on data depth, and the performance of the chart is not reported. There exist few tests for the location parameter of multivariate distribution based on data depth. Using these tests, we proposed nonparametric control charts to detect a shift in the location parameter of the multivariate process. We investigate the performance of the proposed control charts using the average run-length measure for various distributions. Also, the control chart procedure is illustrated by using wine quality data.

Keywords Multivariate processes · Depth functions · DD plot · Bootstrapping

1 Introduction

A multivariate data arise in number of the situations. Most of the multivariate data analysis techniques are based on the assumption of multivariate normality. In this situation, the Hotelling's T^2 chart is widely used to monitor the mean vector of the process. Hotelling's T^2 chart proposed by Hotelling [7] requires multivariate normality of quality characteristics, and justification for multivariate normality is often difficult. In practice, the assumption of multivariate normality may not be fulfilled. Therefore, implementation of Hotelling's T^2 chart is not correct in such situations. Recently, data depth-based nonparametric multivariate analysis techniques have

✉ M. S. Barale
baralemahesh12@gmail.com

D. T. Shirke
dts_stats@unishivaji.ac.in

[1] Department of Statistics, Shivaji University, Kolhapur 416004, India

Published online: 26 March 2019

2 Notion of Data Depth and DD Plot

Let $X \in \mathbb{R}^p$, $p > 1$ be a p variate random variable having distribution $F(\cdot)$ and $X_1, X_2, \ldots, X_n$ be a random sample of size n from $F(\cdot)$. Centrality or outlyingness of a given observation with respect to the given distribution function $F(\cdot)$ or data cloud $X = (X_1, X_2, \ldots, X_n)$ can be measured using the notion of data depth. If $x \in \mathbb{R}^p$, then depth of point x measures how deep/central the point x is with respect to distribution $F(\cdot)$ or data cloud X. Larger is the depth value, deeper is the corresponding observation with respect to the distribution $F(\cdot)$ or data cloud X. This provides a center outward ordering of points which can be useful for multivariate statistical analysis. The depth of any point x; $x \in \mathbb{R}^p$ with respect to distribution $F(\cdot)$ or data cloud X is computed by using the suitable depth function $D_F(x)$. Zuo and Serfling [23] introduced four desirable properties for the depth functions, which are also discussed by Liu [9]. These properties are, namely, (1) affine invariance, (2) maximality at the center, (3) monotonicity relative to the deepest point and (4) vanishing at infinity. One can see more details regarding the depth functions in Zuo and Serfling [23]. There are many depth functions available in the literature such as Mahalanobis depth [14], Tukeys half-space depth [21], Oja depth extension of Oja median [16], projection depth [22] and simplicial depth [9].

A two-dimensional graph known as DD plot is introduced by Liu et al. [11]. The DD plot is very useful diagnostic tool used to compare samples from two multivariate populations. Let F and G be two continuous distribution functions in $\mathbb{R}^p$. Let $X = (X_1, X_2, \ldots, X_n)$ and $Y = (Y_1, Y_2, \ldots, Y_m)$ be the random samples coming from F and G, respectively. Consider the $D_F(x)$ and $D_G(x)$ be the depth values of point x with respect to F and G, respectively.

Consider a set of pairs $DD(F, G) = \{(D_F(x), D_G(x)), \text{for all } x \in \mathbb{R}^p\}$. In practice, the distributions F and G are unknown. Therefore, empirical form of $DD(F, G)$ is considered as $DD(F_n, G_m) = \{(D_{F_n}(x), D_{G_m}(x)), \text{ for all } x \in \{X \cup Y\}\}$. The two-dimensional plot of $DD(F_n, G_m)$ is referred to as DD plot.

Consider the situation where two populations have different location parameters. Panels a, b and c in Figs. 1, 2, 3 and 4 show various patterns of DD plots for various types of shifts in location for Mahalanobis depth, half-space depth and projection depth, respectively. Here, F is taken as bivariate normal with mean $\mu = (0, 0)^T$ and G is bivariate normal with shifted mean $\mu_1 = \mu + \delta 1$, $\delta = 0, 0.5, 1$ and 1.5 with equal scatter. When F and G are identical, then DD plot shows all points clustered along the 45° diagonal line, whatever be the depth function used. If there is any difference in two distributions, points will deviate from the diagonal. DD plot exhibits a noticeable departure from the diagonal line in a symmetric manner. When F is not identical to G, different depth functions show different DD plot patterns, a leaf-shaped pattern for Mahalanobis depth and half-space depth, the star-shaped pattern for projection depth. But the departure from diagonal line pulls down from the point $(\max_t D_{F_n}(t), \max_t D_{G_m}(t))$ to (0, 0), leaving the upper right corner empty and spreading out the points around the midrange of the diagonal line. As shift increases, the points get concentrate

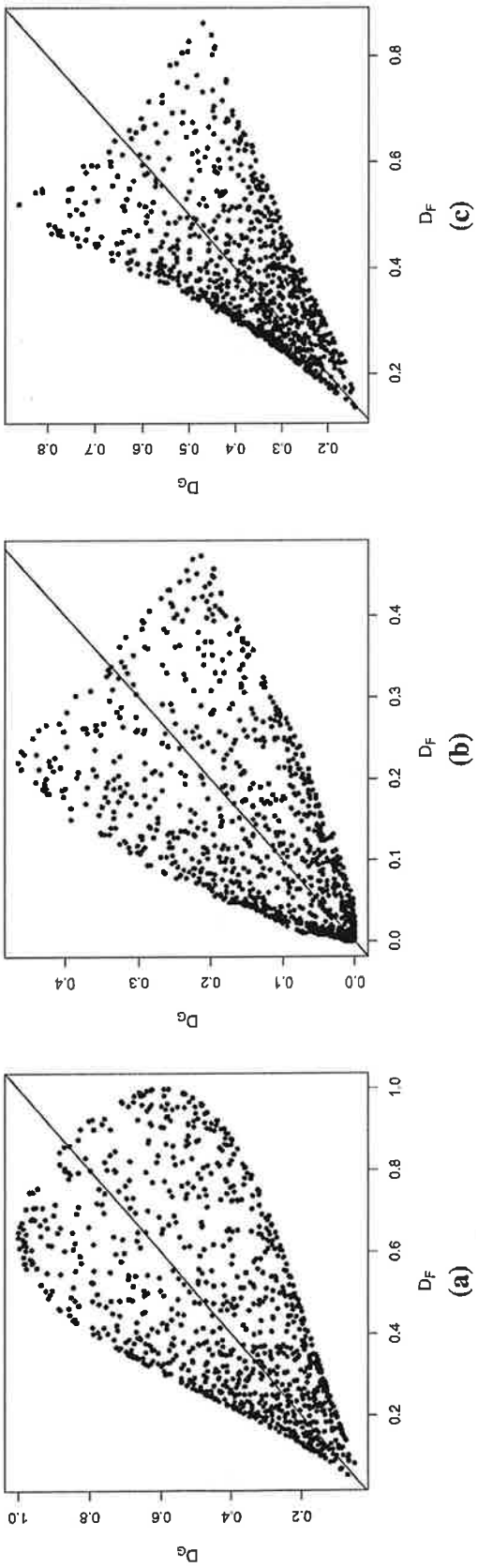

Fig. 2 DD plot when G is $N_2((0.5, 0.5); I_2)$ using Mahalanobis (a), half-space (b) and projection depth (c)

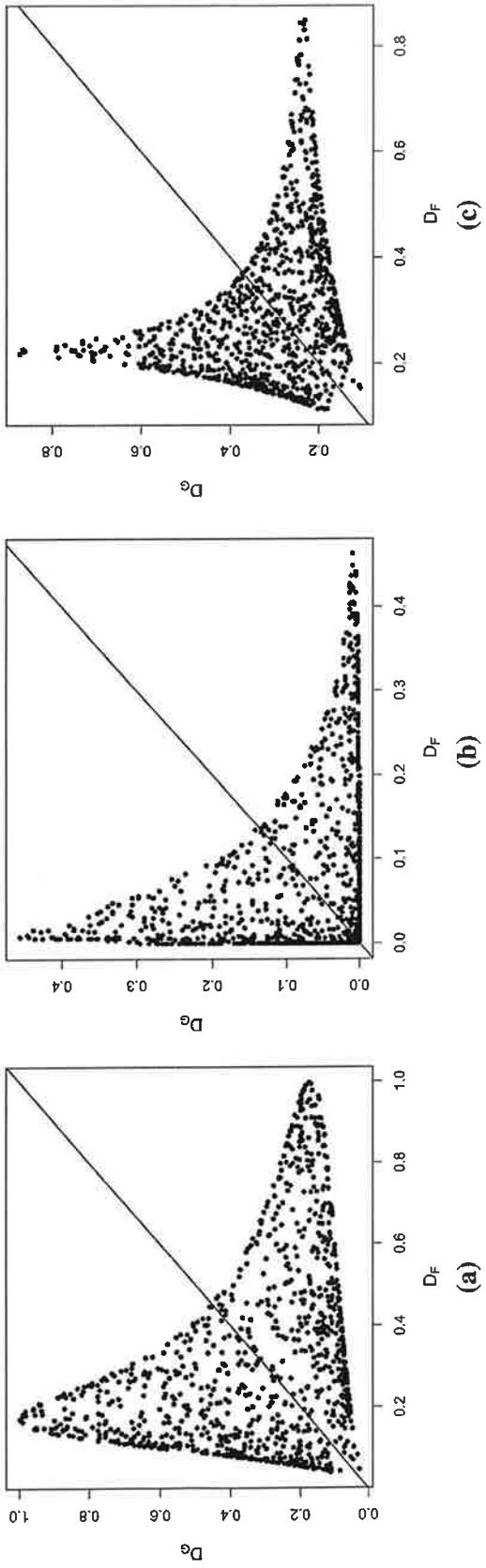

Fig. 4 DD plot when G is $N_2((1.5, 1.5); I_2)$ using Mahalanobis (a), half-space (b) and projection depth (c)

testing the aforementioned null hypothesis. The smaller T_n provide evidence against the null hypothesis, that is, two samples have different location vectors. The p value for the test can be obtained using popular permutation test procedure, and it is given by

$$p_{n,B}^{T} = 1/B \sum_{i=1}^{B} I(T_i^s \leq T_{\text{Obs}}), \tag{1}$$

where $I(a \leq b) = 1$, if $a \leq b$; $= 0$, otherwise T_{Obs} is the observed value of statistic T_n calculated from the combined sample, T_i^s are the values of the statistic T_n computed using ith permuted combined sample $i = 1, 2, \ldots, B$ and B is the total number of permutations.

3.2 W_1-Based and W_2-Based Test

Shirke and Khorate [20] have proposed two nonparametric tests based on data depth differences for testing equality of mean vectors of two multivariate populations. Test statistic provided by Shirke and Khorate [20] are given as follows:

$$W_1 = \frac{1}{m+n} \sum_{i=1}^{m+n} |D_{F_n}(z_i) - D_{G_m}(z_i)|; \qquad z_i \in X \cup Y \tag{2}$$

and

$$W_2 = \frac{1}{m+n} \sum_{i=1}^{m+n} (D_{F_n}(z_i) - D_{G_m}(z_i))^2; \qquad z_i \in X \cup Y. \tag{3}$$

Reject the null hypothesis for larger values of W_1 and W_2. Larger the value of W_1 and W_2 indicate stronger evidence against location shift between the two distributions. The p value is obtained by using Fishers permutation test as follows

$$p_{n,}^{W_1} = 1/B \sum_{i=1}^{B} I(W_{1_i}^s \geq W_{1_{\text{Obs}}}), \tag{4}$$

where $I(a \geq b) = 1$, if $a \leq b$; $= 0$, otherwise $W_{1_{\text{Obs}}}$ is the value of test statistic W_1 calculated from the original combined sample, $W_{1_i}^s$ are the values of the test statistic W_1 computed from ith permuted combined sample $i = 1, 2, \ldots, B$ and B is the number of permutations. On the similar lines, p value for W_2-based test can be obtained. General theory says that for the permutation test, the true p value can be computed by using all possible permutations $\frac{(m+n)!}{m!n!}$ of combined sample. However, Boos and Zhang [3] and Marozzi [15] suggested to use $B = 8\sqrt{M}$, where M is the number of Monte Carlo simulations. Therefore, for $M = 5000, 2000$ and 1000, values of B are choosen to be 566, 358 and 253.

The larger location shift between the two distributions, the smaller is the value of T and the larger values of W_1 and W_2. Thus, the stronger the evidence of the process is out of control. Here, it is difficult to obtain distribution of T, W_1 and W_2. We call these charts as T chart, W_1 chart and W_2 chart.

4.2 Control Limits of T Chart, W_1 Chart and W_2 Chart

Similar to the various multivariate control charts, T chart, W_1 chart and W_2 chart have only single control limit lower control limit (LCL) for T chart and upper control limit (UCL) for W_1 and W_2 chart, respectively. It is not easy to find exact distribution of T, W_1 and W_2. We use the method of bootstrapping to obtain control limits of proposed charts. We give an algorithm for obtaining control limits of the proposed chart as follows:

1. Generate a bootstrap sample of size m from the reference sample of size n.
2. Compute the test statistic with respect to the whole reference sample.
3. Repeat the procedure B time to get B bootstrap sample statistics $(T_1^*, T_2^*, \ldots, T_B^*)$.
4. Compute lower $(100 \times \alpha)$th quantile of $(T_1^*, T_2^*, \ldots, T_B^*)$ as a LCL for T chart.

Similarly, upper $(100 \times (1\text{-}\alpha))$th quantile of W_{1i}^* (W_{2i}^*); $i = 1, 2, \ldots B$ is used as UCL for W_1 (W_2) chart. Generally, B should be sufficiently large enough. In practice, $B = 1500$ to 2000 is sufficient. The proposed charts implement data-dependent control limits, which will vary as reference sample size and subgroup sample size vary. A usual Shewhart-type control chart uses fixed control limits based on distribution of charting statistics. In this scenario, the control limits vary even if data follow an identical in-control distribution, which does not require information about in-control distribution F. While using bootstrapping there is no need of actual in-control distribution F, the limits obtained might not be accurate to study run-length distribution properties when the reference sample size n is too small.

5 Performance Study

An average run length (ARL) is one of the performance measures used for comparing the control charts. ARL is defined as the expected number of samples required to get a first out-of-control signal, and it can be obtained by taking reciprocal of false alarm probability. A control chart with minimum out-of-control ARLs shows better assignable cause detection ability. The proposed charts are compared by taking in-control ARL ≈ 200 with the Q chart due to Liu [10] based on data depth by using R software. We have considered bivariate normal, bivariate Cauchy and bivariate skew normal distribution as in-control process distributions, and out-of-control observations are generated by giving shift δ in location parameter of in-control process distributions. Here, we have taken reference sample of size $n = 100$ primarily and subgroup sample of size $m = 10, 15$ for various shifts $\delta = 0.2, 0.4, 0.6, 0.8$ and 1 in location parameter. The control limit is obtained by using the method of

Table 1 Comparison of Q, T, W_1 and W_2 charts for different depth function when $n = 100$

Depth function		Half-space				Projection			
m	δ	Q chart	T chart	W_1 chart	W_2 chart	Q chart	T chart	W_1 chart	W_2 chart
Normal									
10	0	202.84	205.80	209.65	216.48	208.01	200.27	202.31	200.00
	0.2	139.64	123.69	100.14	85.00	142.67	120.49	168.28	155.86
	0.4	61.58	38.63	20.09	20.45	54.88	37.85	106.03	88.98
	0.6	22.26	11.30	5.82	5.85	20.09	12.13	54.63	42.37
	0.8	7.06	3.99	2.38	2.44	6.77	4.68	27.51	22.13
	1	2.92	2.01	1.44	1.49	2.81	2.36	15.59	11.95
15	0	209.09	216.98	206.60	201.04	206.56	202.66	208.15	213.60
	0.2	143.65	124.69	74.74	76.52	125.17	110.10	156.54	148.99
	0.4	49.10	31.22	11.87	13.09	48.39	25.89	81.89	66.50
	0.6	14.47	6.88	3.05	3.37	13.09	6.96	35.69	25.44
	0.8	4.59	2.34	1.49	1.55	4.26	2.65	15.40	11.08
	1	1.88	1.34	1.10	1.14	1.90	1.50	7.36	5.95
Cauchy									
10	0	200.97	201.95	206.08	200.24	199.53	201.45	204.04	200.93
	0.2	200.92	123.76	144.03	137.21	177.14	158.08	173.46	170.71
	0.4	176.70	54.24	64.76	55.99	159.16	73.20	117.40	96.62
	0.6	137.82	22.37	26.62	22.90	117.13	29.56	72.31	52.44
	0.8	93.45	9.01	12.79	10.57	72.33	11.15	40.60	26.49
	1	68.71	185.29	7.03	5.59	43.89	4.79	23.93	15.47
15	0	199.25	203.38	208.02	200.98	201.74	202.26	211.02	208.67
	0.2	180.42	169.60	138.00	123.81	197.08	132.92	169.70	155.62
	0.4	137.34	81.29	50.47	45.07	147.09	50.13	101.64	66.07
	0.6	101.27	31.43	20.16	16.90	91.17	15.73	53.63	31.01
	0.8	68.57	10.18	9.83	8.77	58.54	5.34	25.79	14.35
	1	47.00	3.97	5.65	5.19	27.04	2.44	13.73	7.46
Skew normal									
10	0	215.91	203.20	206.02	200.49	201.48	202.60	205.92	202.30
	0.2	580.78	89.18	47.17	57.07	92.15	67.44	136.85	119.82
	0.4	142.62	13.96	4.97	4.51	19.18	12.49	53.64	44.61
	0.6	16.60	2.64	1.36	1.24	3.78	3.08	19.79	15.27
	0.8	3.12	1.15	1.01	1.01	1.35	1.35	7.88	6.29
	1	1.25	1.01	1.00	1.00	1.02	1.03	3.93	3.19
15	0	205.52	205.19	208.07	208.60	199.52	201.38	210.32	208.03
	0.2	721.50	73.30	25.08	28.39	88.83	58.62	122.05	102.26
	0.4	141.77	8.27	1.96	2.00	14.47	7.73	36.70	26.84
	0.6	10.61	1.55	1.03	1.01	2.46	1.86	11.12	7.60
	0.8	1.82	1.02	1.00	1.00	1.10	1.08	3.94	3.21
	1	1.04	1.00	1.00	1.00	1.00	1.00	2.03	1.72

$\gamma = 1.5, 2$ to the scale matrix. Table 3 describes the ARL values when there is a shift in location and scale both. It is observed that the result described above holds. The T

Table 3 Comparison of Q, T, W_1 and W_2 charts along with shift in scatter matrix for $n = 100$ and $m = 10$

δ	γ	Q chart	T chart	W_1 chart	W_2 chart
Normal					
0	1	202.21	202.13	200.29	205.71
0.5	1	36.92	20.66	76.57	61.44
1	1	2.89	2.35	14.98	12.04
1.5	1	1.09	1.12	4.87	4.08
0	1.5	205.01	95.37	315.46	265.93
0.5	1.5	38.45	16.99	119.47	93.28
1	1.5	2.94	2.84	24.99	19.55
1.5	1.5	1.10	1.27	7.85	6.58
0	2	207.98	57.38	357.93	284.07
0.5	2	35.49	15.29	146.94	108.13
1	2	2.97	3.23	34.15	26.01
1.5	2	1.09	1.41	11.96	9.34
Cauchy					
0	1	199.65	204.94	203.02	206.65
0.5	1	137.61	45.15	95.06	73.31
1	1	45.27	4.72	23.31	16.04
1.5	1	13.26	1.54	7.88	5.27
0	1.5	202.38	101.09	286.97	264.16
0.5	1.5	132.37	31.69	153.24	112.68
1	1.5	49.30	5.33	42.93	26.00
1.5	1.5	13.34	1.79	14.00	8.79
0	2	187.70	69.24	301.95	277.35
0.5	2	129.61	26.41	183.91	130.83
1	2	47.39	5.43	57.47	35.52
1.5	2	13.14	2.08	19.46	12.38
Skew Normal					
0	1	205.98	201.09	205.18	200.34
0.5	1	8.30	5.51	32.62	24.39
1	1	1.02	1.03	3.97	3.25
1.5	1	1.00	1.00	1.50	1.41
0	1.5	205.23	70.95	269.00	229.60
0.5	1.5	8.08	3.56	30.25	23.40
1	1.5	1.02	1.03	4.88	4.10
1.5	1.5	1.00	1.00	1.88	1.74
0	2	191.30	31.44	240.84	184.75
0.5	2	8.52	2.68	28.47	21.98
1	2	1.01	1.04	5.62	4.88
1.5	2	1.00	1.00	2.23	2.06

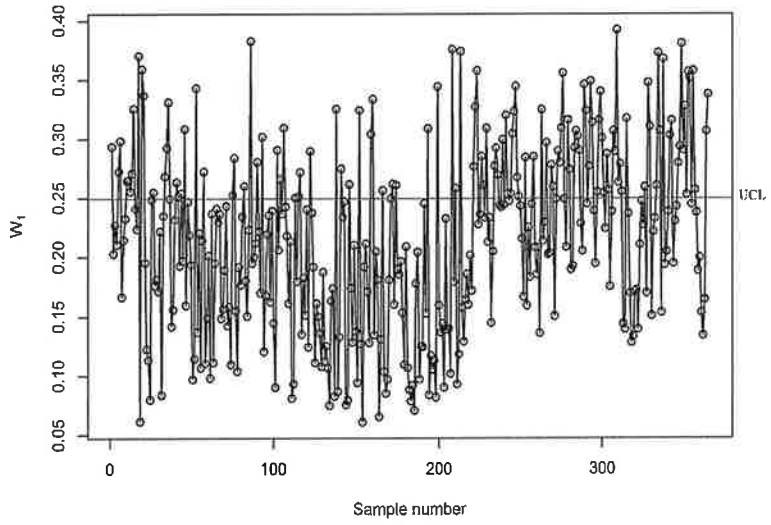

Fig. 6 W_1 chart for wine quality data

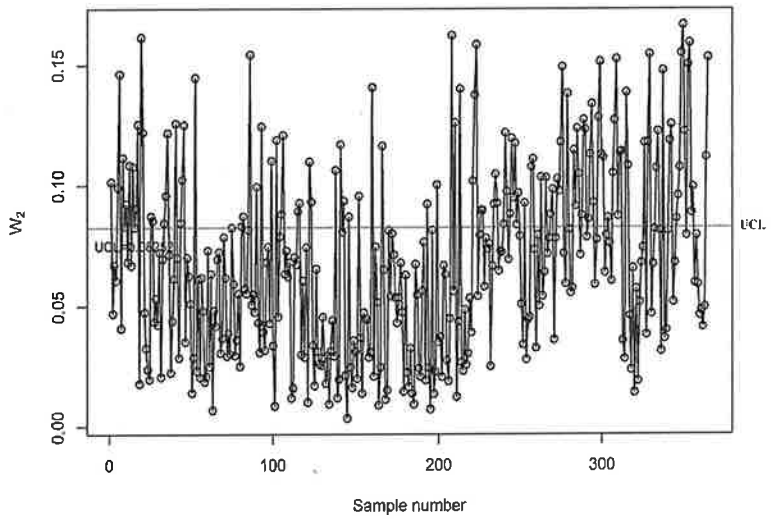

Fig. 7 W_2 chart for wine quality data

correctly of quality level "six" and 118 out of 145 samples of quality level "five". A W_1 chart detects 50 samples as an out-of-control out of 219 samples and 74 samples out of 145. A W_2 chart detects 48 samples as an out-of-control out of 219 samples and 73 samples out of 145. It can be seen that T chart has less type II error as compared to W_1 and W_2 charts.

7 Concluding Remarks

There are very few methods are available to tackle the problem of multivariate non-normality in the area of multivariate statistical process control. In the present paper, we provide the control charts based on the notion of data depth. A notion

13. Lowry CA, Montgomery DC (1995) A review of multivariate control charts. IIE Trans 27(6):800–810
14. Mahalanobis PC (1936) On the generalized distance in statistics. National Institute of Science of India, Calcutta
15. Marozzi M (2016) Multivariate tests based on interpoint distances with application to magnetic resonance imaging. Stat Methods Med Res 25(6):2593–2610
16. Oja H (1983) Descriptive statistics for multivariate distributions. Stat Probab Lett 1(6):327–332
17. Osei-Aning R, Abbasi SA, Riaz M (2017) Bivariate dispersion control charts for monitoring non-normal processes. Qual Reliab Eng Int 33(3):515–529
18. Phaladiganon P, Kim SB, Chen VC, Baek J-G, Park S-K (2011) Bootstrap-based T^2 multivariate control charts. Commun Stat Simul Comput ® 40(5):645–662
19. Polansky AM (2005) A general framework for constructing control charts. Qual Reliab Eng Int 21(6):633–653
20. Shirke D, Khorate S (2017) Power comparison of data depth-based nonparametric tests for testing equality of locations. J Stat Comput Simul 87(8):1489–1497
21. Tukey JW (1975) Mathematics and the picturing of data. In: Proceedings of the international congress of mathematicians, vol 2. Vancouver, pp 523–531
22. Zuo Y et al (2003) Projection-based depth functions and associated medians. Ann Stat 31(5):1460–1490
23. Zuo Y, Serfling R (2000) General notions of statistical depth function. Ann Stat 28:461–482