Extensions for *DD*-plot

Dae-Heung Jang[1]
[1] Pukyong National University, Busan, KOREA dhjang@pknu.ac.kr

Abstracts

The *DD*-plot is a useful graphical exploratory data analysis tool for graphical comparisons of two multivariate distribution or samples based on data depth. We can suggest several extensions for *DD*-plot, namely, *DDD*-plot, *DD*-plot matrix, three-dimensional *DD*-plot, and dynamic *DD*-plot. If we are interested in comparisons of three multivariate distributions *F*, *G*, and *H*, we can suggest *DDD*-plot as an extension of *DD*-plot. If we are interested in comparisons of *q*(>2) multivariate distributions $F_1, F_2, ..., F_q$, we can suggest *DD*-plot matrix as an extension of *DD*-plot. If we are interested in comparisons of two multivariate distributions *F* and *G* with control parameter(eg. time), we can suggest dynamic *DD*-plot and three-dimensional *DD*-plots as an extension of *DD*-plot.

Keywords: data depth, *DD*-plot, *DDD*-plot, dynamic *DD*-plot

1. Introduction

Data depth is a way of measuring how deep(or central) a given point is with respect to a distribution or a given data cloud. There are many examples about data depth(Mahalanobis, half-space, convex hull peeling, Oja, simplicial, majority, likelihood, projection, zonoid, special, spherical, lens, etc.).
The *DD*-plot was supposed by Liu et al.(1999). The *DD*-plot is a useful graphical exploratory data analysis tool for graphical comparisons of two multivariate distribution or samples based on data depth. Numerical examples, Li et al.(2012) showed that *DD*-classifier using the *DD*-plot is comparable or better than k-nearest neighbor or the support vector machine methods and the *DD*-classifier performs well in general settings, including nonelliptical distributions or elliptical distributions with unequal priors and with simultaneous difference in location and scale.

DD-plot is nonparametric, completely data-driven, and simple to visualize and *DD*-plot is easy to implement and robust against outliers and extreme values.

We can suggest several extensions for *DD*-plot, namely, *DDD*-plot, *DD*-plot matrix, three-dimensional *DD*-plot, and dynamic *DD*-plot. If we are interested in comparisons of three multivariate distributions *F*, *G*, and *H*, we can suggest *DDD*-plot as an extension of *DD*-plot. If we are interested in comparisons of *q*(>2) multivariate distributions $F_1, F_2, ..., F_q$, we can suggest *DD*-plot matrix as an extension of *DD*-plot. If we are interested in comparisons of two multivariate distributions *F* and *G* with control parameter(eg. time), we can suggest dynamic *DD*-plot and three-dimensional *DD*-plots as an extension of *DD*-plot.

2. *DDD*-plot and *DD*-plot matrix

Sometimes, we are interested in comparisons of three multivariate distributions F, G, and H(eg. MANOVA). If we are interested in comparisons of three multivariate distributions *F*, *G*, and *H*, we can suggest *DDD*-plot as an extension of *DD*-plot.

Let $\{X_1, X_2, ..., X_l\}(\equiv \mathbf{X})$, $\{Y_1, Y_2, ..., Y_m\}(\equiv \mathbf{Y})$, and $\{Z_1, Z_2, ..., Z_n\}(\equiv \mathbf{Z})$ be three random samples from F, G and H, respectively, which are distributions defined on R^d. If $F = G = H$, the *DDD*-plot should be concentrated along the 45-degree 3-D line. If $\{F = G = H\}$ is not true, the *DDD*-plot would exhibit a noticeable departure from the 45-degree 3-D line.

Figure 1(a) shows *DDD*-plot for Fisher's iris data. We can find three clusters. Three species are *iris setosa*(F), *iris versicolor*(G), *iris virginica*(H). Each species consists of 50 observations and 4 variables(sepal length, sepal width, petal length, petal width). Figure 1(b) shows *DDD*-plot for cereal data(Johnson and Wichern(2007)). Cereal data consists of 43 brands(General Mills(F, 17 brands), Kellogg(G, 20 brands), Quaker(H, 6 brands) and 8 variables(calories, protein, fat, sodium, fiber, carbohydrates, sugar, potassium). We can find two clusters(1: General Mills and Kellogg, 2: Quaker).

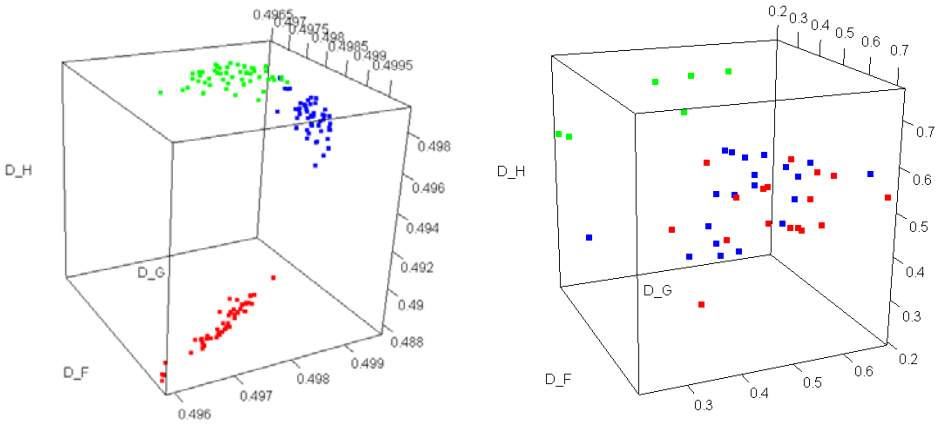

Fig. 1. (a) *DDD*-plot for iris data(red: *iris setosa*, blue: *iris versicolor*, green: *iris virginica*, data depth: Oja Depth) (b) *DDD*-plot for cereal data(red: General Mills, blue: Kellog, green: Quaker, data depth: Modified Band Depth)

For simulation, we suppose that in-control state in multivariate process is $N_4(\boldsymbol{\mu}, \Sigma)$ with $\boldsymbol{\mu} = (0,0,0,0)$

and $\Sigma = \begin{pmatrix} 1.000 & -0.424 & 0.957 & 0.922 \\ -0.424 & 1.000 & -0.540 & -0.063 \\ 0.957 & -0.540 & 1.000 & 0.822 \\ 0.922 & -0.063 & 0.822 & 1.000 \end{pmatrix}$

and that there are 3 production lines. Assume that Production line 1(F) keeps in-control state and that there are the change of μ_1 from 0 to 2 in production line 2(G) and the change σ_3^2 from 1 to 9 in production line 3(H). Figure 2 shows *DDD*-plot for simulation data. We can find a considerable change in *DDD*-plot according to the out-of-control state.

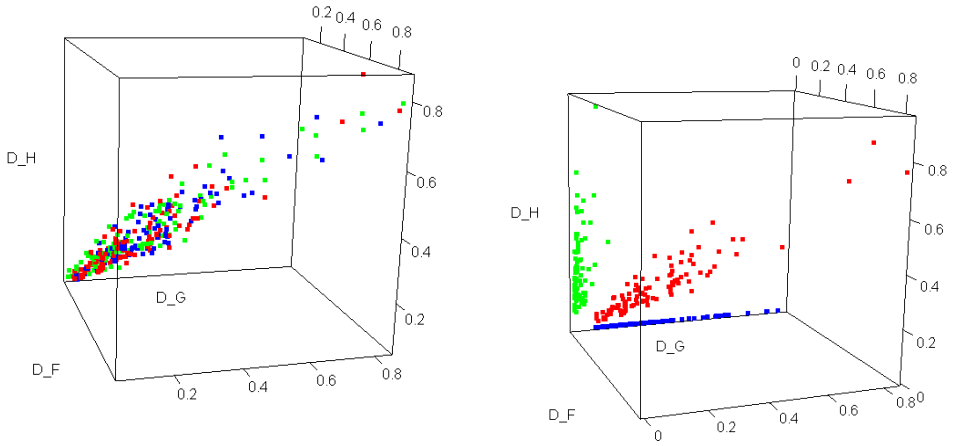

Fig. 2. *DDD*-plot for simulation data(production line 1(red), production line 2(blue), production line 3(green), data depth: Mahalanobis Depth)
(a) In-control state (b) Out-of-control state

Sometimes, we are interested in comparisons of q (>2) multivariate distributions $F_1, F_2, ..., F_q$. We can suggest *DD*-plot matrix as an extension of *DD*-plot.

Figure 3 Shows *DD*-plot matrix for forensic glass fragments data(Venable and Ripley(2002)). Forensic glass fragments data consists of total 205 glass fragments, 5 different types(window float glass(F_1, 70 glass fragments), window non-float glass(F_2, 76 glass fragments), vehicle window glass(F_3, 17 glass fragments), containers(F_4, 13 glass fragments), vehicle headlamps(F_5, 29 glass fragments)) and 9 measured physical characteristics(refractive index, Na(sodium), Mg(Manganese), Al(Aluminium), Si(Silicon), K(Potassium), Ca(Calcium), Ba(Barium), Fe(Iron)).

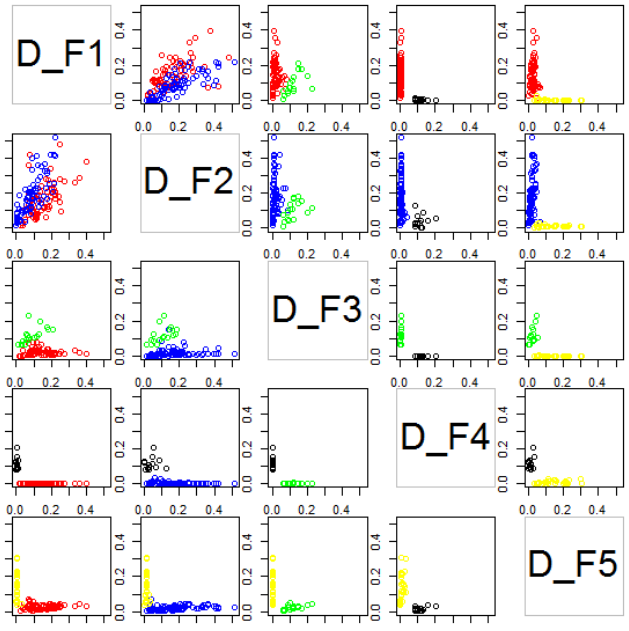

Fig. 3. *DD*-plot Matrix for forensic glass fragments data(Red: window float glass, Blue: window non-float glass, Green: vehicle window glass, Black: containers, Yellow: vehicle headlamps, Data Depth: Mahalanobis Depth)

3. Dynamic *DD*-plot and Three-dimensional *DD*-plot

Sometimes, we are interested in comparisons of two multivariate distributions F and G with control variable(eg. time). We can suggest dynamic *DD*-plot and three-dimensional *DD*-plots as an extension of *DD*-plot.

Reconsider the simulation data in section 2 and consider the following three scenarios.

Scenario 1(S1): The value of μ_1 change slowly from 0 to 2.

Scenario 2(S2): $\boldsymbol{\mu}$ change slowly from $\boldsymbol{\mu} = (2,0,0,0)$ to $\boldsymbol{\mu} = (2,0,0,3)$.

Scenario 3(S3): The value of σ_3^2 change slowly from 1 to 9.

Figure 4 shows dynamic DD-plot and quality index plot for the out-of-control state $\mu_1 = 0.35$ in a multivariate process using Oja depth in Scenario 1(S1).

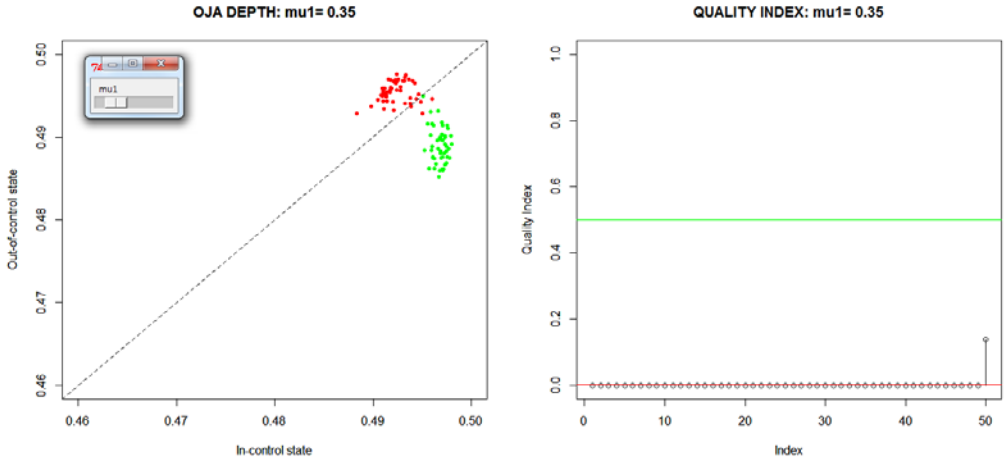

Fig. 4. Dynamic DD-plot and quality index plot for the out-of-control state $\mu_1 = 0.35$ in a multivariate process using Oja depth in Scenario 1(S1)

Figure 5 shows Three-dimensional DD-plot for the out-of-control state with the change of μ_1 from 0 to 2 in a multivariate process using Oja depth in Scenario 1(S1).

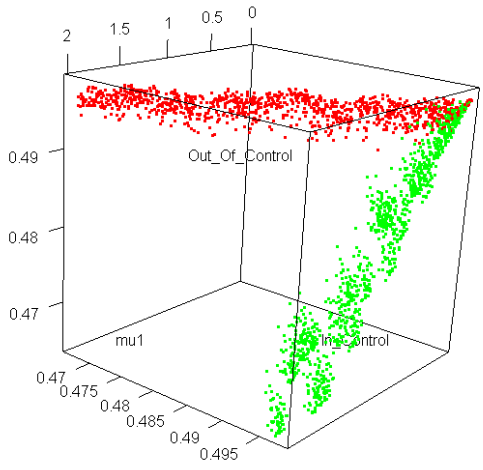

Fig. 5. Three-dimensional DD-plot for the out-of-control state with the change of μ_1 from 0 to 2 in a multivariate process using Oja depth in Scenario 1(S1)

Figure 6 shows dynamic DD-plot and quality index plot for the out-of-control state $\sigma_3^2 = 9$ in a multivariate process using Mahalanobis depth in Scenario 3(S3).

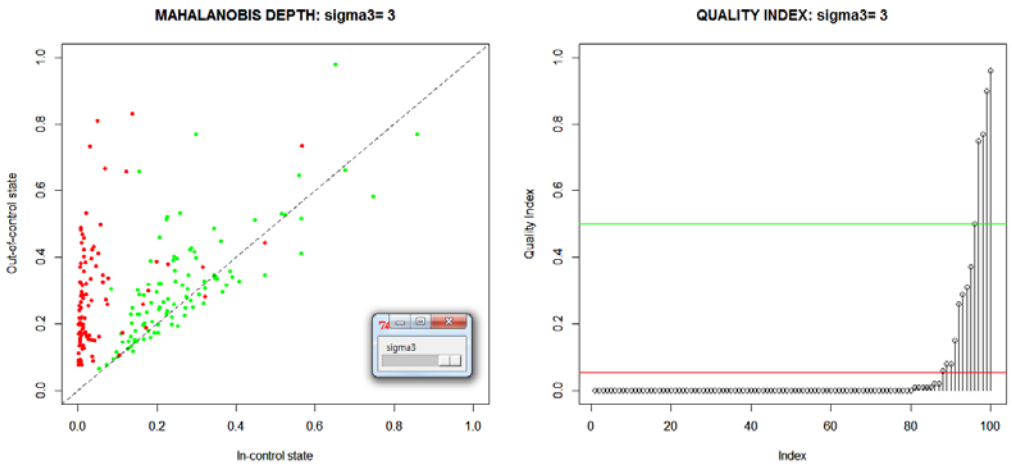

Fig. 6. Dynamic DD-plot and quality index plot for the out-of-control state $\sigma_3^2 = 9$ in a multivariate process using Mahalanobis depth in Scenario 3(S3)

4. Conclusions

As extensions of *DD*-plot: *DDD*-plot, we can suggest *DD*-plot matrix, Dynamic *DD*-plot, and Three-dimensional *DD*-plot. We can use these extensions of *DD*-plot as the additional graphical tools for *DD*-plot.

References

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, 6th ed., Pearson, New York.

Li, J., Cuesta-Albertos, J. A. and Liu, R. (2012). *DD*-Classier: Nonparametric classication procedure based on DD-plot, *Journal of the American Statistical Association*, **107**, 737-753.

Liu, R., Parelius, J. M. and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference, *The Annals of Statistics*, **27**, 783-858.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S-Plus*, Springer, New York.