TEST Engineering & Management

March-April 2020
ISSN: 0193-4120 Page No. 1273 - 1277

Control Charts for Multivariate Non Normal Data Using Winsorization with Comedian when Outliers are Present in Training Data

[*1]Latha V, [2]P Rajalakshmi

[1]Associate Professor, Department of Statistics, Jyoti Nivas College, Autonomous, Bangalore
[2]Professor (Retd.), Department of Statistics, Bangalore University, Bangalore
[1]latha.v.1968@gmail.com

Article Info
Volume 83
Page Number: 1273 - 1277
Publication Issue:
March - April 2020

Article History
***Article Received**: 24 July 2019*
***Revised**: 12 September 2019*
***Accepted**: 15 February 2020*
***Publication**: 14 March 2020*

Abstract

Robust control charts have been proposed for multivariate data as the traditional control charts are affected by extreme observations. Control charts have also been constructed for multivariate non normal data. Our proposed control chart uses comedian as a outlier detection measure along with winsorization to construct control limits in phase I of the chart so that most of the information that is available is utilized and not discarded because of outlier observations. Simulation and winsorization have been used to generate observations from multivariate skew t and the g and h distributions.

***Keywords:** Non-normal data, winsorization, comedian, skewness, g and h distribution, multivariate skew distribution*

1. Introduction

Control charts are statistical tools for monitoring various industrial processes as indicators of change in shift in the process characteristics. The statistical process control chart was first developed by W.A Shewhart in 1931 for monitoring univariate statistical process control. Univariate control charts can monitor only one variable at a time and cannot study the relationship between the several variables which are usually present in a manufacturing process. The assumption of the process values following normal distribution and being independent and identically distributed may not always be valid given the dynamic behavior of the variables. Most of the manufacturing processes are affected by disturbances known and unknown. Individual monitoring of process variables ignores the correlation and interaction between variables. To include the correlation between variables and to study the relationship between them, multivariate control charts were developed in the 1940’s by Hotelling(1947). Some of the multivariate statistical process control charts are HotellingT^2, T^2 generalized variance charts and Exponentially Weighted Moving Average(EWMA) charts.

A multivariate control chart can be used as a tool for detecting shifts in the mean, distributional deviations from the in control distribution and for detecting multivariate outliers. An overview of the developments in multivariate statistical process control and its use in fault detection and isolation can be found in Kourti(2005). The multivariate EWMA chart for unequal sample sizes was proposed by Kim and Reynold(2005). The problem of monitoring variance shifts in a multivariate time series was studied by Chang and Zhang(2007) who proposed the MEWMV charts. A single chart which used the EWMA procedure and generalised LR test for joint monitoring of process mean and variability was proposed by Zhang and Chang(2008). Zou and Qui(2009) proposed multivariate statistical process control using LASSO. A multivariate logistic regression model was proposed by Sinha S.K et al.(2010) for analysis of multiple binary outcomes with incomplete covariate data.

Traditional control based Principal Component Analysis control charts was first proposed by Jackson(1999). Q charts based on the residuals of the principal components was proposed by Ferrer(2007). Principal Component Analysis based control charts for multivariate non normal data was proposed by Phaladiganon P et al. (2013). They proposed two methods using PCA with kernel density estimation and PCA with bootstrapping. Ahsan et al. (2018) proposed a multivariate control chart based on PCA mix using kernel density estimation for mixed data by using principal component analysis for continuous variables and multiple correspondence analysis for categorical variables. PCA mix proposed by Chavent et al. (2014) analyses different types of quality characteristics together.

Robust control charts for bivariate data as alternatives to Hotelling'sT^2 chart have been proposed by Abu Shawiesh et al.(2012) using median, median absolute deviation and comedian as measures of location, scatter and covariation respectively. They also conducted a comparison (2014) of bivariate robust control charts using Hotelling'sT^2, T^2_{MEDMAD}, T^2_{MVE}, T^2_{MCD} for individual observations. Farokhnia and Niaki(2019) have proposed a PCA based chart using support vector machines for multivariate non normal distributions. Ahsan et al. (2019) have proposed a control chart for outlier detection using PCA mix for mixed data.

Our proposed control chart uses comedian as a outlier detection measure along with winsorization to construct control limits in phase I of the chart so that most of the information that is available is utilized and not discarded because of outlier observations. Simulation is done by generating samples from.

i. Standard normal distribution with $\mathbf{N_p(0, I_p)}$

ii. Multivariate skew t distribution: Let Y be a p-dimensional random vector. Then Y is said to follow a p-dimensional unrestricted skew t distribution with p × 1 location vector μ, p × p scale matrix Σ, p × 1 skewness vector δ, and (scalar) degrees of freedom ν, if its probability density function is given by

$f_p(y; \mu, \Sigma, \delta, \nu) = 2^p t_{p,\nu}(y; \mu, \Omega)\ T_{p,\nu+p}(y^* ; 0, \Lambda)$,

Where $\Delta = \text{diag}(\delta)$, $\Omega = \Sigma + \Delta^2$, $y^* = q\ \sqrt{}(\nu + p/(\nu + d(y))$,

$q = \Delta\Omega^{-1} (y - \mu)$, $d(y) = (y - \mu)\ `\Omega^{-1} (y - \mu)$, $\Lambda = I_p - \Delta\Omega^{-1}\Delta$.

iii. g x h distribution: The g x h distribution is used when skewness and elongation are jointly considered. g measures the skewness and h is a measure of elongation. It is given by

$Y_{g,h}(Z) =(e^{gz}-1)g^{-1}e^{\frac{hz2}{2}}$; Z is a standard normal variate.

2. Traditional Hotelling T^{2}Chart

Two phases are described in constructing multivariate control charts. In phase I, the parameters of the in control process are estimated and control limits are set up using training data or historical data which is obtained when the process is inthe normal region. In phase II, the estimates and control limits in phase I are used to test whether the manufacturing process is in control or not.

Multivariate statistical process control methods were first developed as an extension to univariate Shewhart control charts, by Hotelling(1947) using the T^2 statistic. Let $X=(x_1,x_2,x_3,\ldots x_n)$ be a sample where x_i is a p dimensional vector of measurements made at time period i. Then Hotelling T^2 statistic is

$T_i^2= (x_i-\mu)'\Sigma^{-1}(x_i-\mu)$

When the process is in control, it is assumed that x_i's are independent and follow multivariate normal distribution N(μ,Σ) where μ is the mean vector and Σ is the variance covariance matrix. When μ and Σ are known, the statistic follows Chi square distribution with p degrees of freedom.

When μ and Σ are unknown, they are estimated using the sample mean vector and the sample covariance matrix respectively. Then $T_i^2= (x_i-x)'S^{-1}(x_i-x)$, where x and S are the location and scale estimates. When T^2 is computed under the null hypothesis, further inference about the process can be made by comparing it to the defined limits which are obtained using $T^2_{CL} = p(n+1)(n-1)F_{\alpha,p,n-p}/n^2$ where n denotes the number of sample size and p denotes the number of quality characteristics. These estimates are sensitive to outliers and hence the T^2 statistic is also sensitive to outliers.

3. Outlier Detection Techniques and Winsorization

It is well known that outliers affect the process of estimation and inference. The measures of location and scatter are unduly affected by outliers. Hence there is a need to study their influence and identify methods to reduce their effect or eliminate them from the data sets. This is a challenging aspect of data analysis. Various methods and techniques have been devised to detect multivariate outliers. Some of these methods are distance based. Mahalanobis squared distance is one such measure used for multivariate data analysis. A large value of this measure indicates that the corresponding observation is an outlier. The problem of 'masking' and 'swamping' also exist in the sense that there are outliers who have a very small value for the Mahalanobis distance and a large value of the distance measure need not necessarily indicate an outlier. Therefore there is a need to tackle this problem by using robust distances which are obtained by replacing the classic estimates by estimators which are robust. Some of the robust estimators proposed are the affine equivariant M estimators proposed by Maronna(1976), StahelDonoho estimators(1982) which are the weighted mean vector and the covariance matric with weights depending on the outlyingness of an observation, Minimum covariance determinant estimator by Rosseeuw(1984), a fast outlier detection procedure proposed by Pena and Prieto(2001) using the direction of the projections that maximise and minimize the coefficient of kurtosis of the projected data and the orthogonalised Gnanadesikan-Kettenring estimator proposed by Maronna and Zamar(2002) to obtain affine equivariant robust scatter matrices beginning with any pairwise robust scatter matrix, which performs well under high collinearity.

A method to detect multivariate outliers has been proposed by Sajesh and Srinivasan(2012) using the measure Comedian defined by Falk(1997). This method can detect a large number of outliers. Falk introduced a dependence measure called the comedian which is a robust measure of the covariance between random variables U and V. For any two random variables U and V, comedian is defined as,

COM(U,V)= med((U-med(U))(V-med(V)))

Where med(U) and med(V) are the medians of U and V respectively. It is equal to the square of the median absolute deviation (MAD) when U=V and has the highest breakdown point. Further COM(U,V) always exists, is symmetric and location and scale invariant. An alternative to the coefficient of correlation based on the median is called the correlation median and is given by

δ(U,V)= COM(U,V)/(MAD(U)MAD(V)

Where MAD(U) and MAD(V) are the median absolute deviations of U and V respectively. Falk proposed this measure as a measure of dependence.

The method proposed by Sajesh and Srinivasan for outlier detection is as follows-

Let U=($\mathbf{u}_{ij}$) , i=1,2…n; j=1,2…p be a n x p matrix.

The comedian matrix is COM(U) =COM($\mathbf{u_i}$,$\mathbf{u_j}$), i,j=1,2,…p and the multivariate correlation median matrix is δ(U)= DCOM(U)D' where D is a diagonal matrix with diagonal elements being the reciprocal of MAD($\mathbf{u}_i$), i=1,2...p. As the comedian matrix is not positive definite, the following steps are used so that the estimators obtained are robust.

The eigen values λ_i and eigen vectors e_j of the scatter matrix such that δ(U)=EΛE' where E is the matrix with columns as ej's and Λ=diag(λ_1, λ_2,… λp). Then define Q=D(U)$^{-1}$E where D is as defined above. Let $\mathbf{z}_i$= Q^{-1}u$_i$, i=1,2…n where z$_i$' is the ith row of the orthogonal matrix Z. Then the robust location and scatter estimates are

m(U)= QF where F=(med($\mathbf{z}_1$),med($\mathbf{z}_2$),…med($\mathbf{z}_j$))

S(U)= QΓQ' where Γ=diag(t_1^2, t_2^2,…t_p^2) where t_j= MAD($\mathbf{z_j}$), j=1,2,…p

The robust Mahalanobis distance is given by

RD(u$_i$, m)= rd$_i$= (u$_i$-m)'S^{-1}(u$_i$-m) where m and S are as defined above.

The cutoff value for identifying outliers is given by

$$C= \frac{1.4826\,(\chi^2_{p\,(0.95)})}{\chi^2_{p\,(0.5)}} med(rd_1, rd_2, \ldots rd_n)$$

If any RD(u$_i$, m)>C, then u$_i$ is an outlier. The expression for C is obtained following Maronna and Zamar(2002) and holds for non-normal original data.

Estimation of parameters is done using the trimmed values obtained after the winsorization process. The given observations are arranged in ascending order to identify outliers and the outliers are replaced by the corresponding trimmed values. This is done by replacing the outliers less than the smallest value by the smallest value retained and the outliers greater than the largest value by the largest value retained. Hence the winsorized sample is given by

$$W_{ij}=\begin{cases} u_{(i_1+1)j}, & if \;\; u_{ij} \le u_{(i_1+1)j}, \\ uij & if \;\; u_{(i+1)j}, \le u_{ij} \le u_{(n-i_2)j}, \\ u_{(n_j-i_2)j}, & if \; u_{ij} \ge u_{(n-i_2)j}, \end{cases}$$

i=1,2..n; j=1,2…p

Then the estimate of the winsorized location measure is m(W)= QF where F=(med($\mathbf{w}_1$),med($\mathbf{w}_2$),…med($\mathbf{w}_p$)) and the scatter estimate is given by S(W)= QΓQ' where Γ=diag(t_1^2, t_2^2,…t_p^2) where t_j= MAD($\mathbf{w}_j$), j=1,2,…p

The distribution of Hotelling's T^2 is unknown under the assumption of non-normal data. In Phase I, the upper control limit for the proposed control chart is computed using simulation and bootstrapping using the overall false alarm rate as α.

4. Simulation and Bootstrapping

Bootstrapping, Efron and Tibshirani(1986), is a resampling procedure which is not based on any assumptions about the parent distribution. It is very useful when the distribution under consideration is non-normal. If x_1,x_2,x_3,…x_n is a random sample of size n then the bootstrapping procedure generates B samples with replacement of size n from the original sample. To construct control limits we use the bootstrapping technique and draw B samples of size n from an in control data. For each bootstrap sample the 100(1-α) percentile value is computed, 0<α<1. An average of the B percentile values is taken as the UCL. If a value exceeds this UCL it is said to be out of control.

Taking α as 0.01, simulation of 5000 data sets of size n from

iv. Standard normal distribution with $\mathbf{N_p(0, I_p)}$
v. Multivariate skew t distribution
vi. g x h distribution

Traditional and robust estimators are then computed.

Phase I involves simulation of 5000 data sets from $\mathbf{N_p(0, I_p)}$ with α = 0.01 and computing the robust estimators. Phase II involves generating an additional observation for each data set and computing the robust statistic using the corresponding estimators obtained in Phase I. The UCL for the robust statistic is obtained by finding the 99th percentile for each sample and averaging it using the median.

For independent variables, we use the mixture normal distribution as

(1-ε) $\mathbf{N_p(0, I_p)}$ +ε $\mathbf{N_p(\mu_1, I_p)}$ where ε is the proportion of outlier data, 0 is the in control mean vector , $\boldsymbol{\mu_1}$ is the out of control mean vector (taking values 0, 2, 5) and I_p is the identity dispersion matrix. For dependent variables, the mixture normal distribution is

(1-ε) $\mathbf{N_p(0, \Sigma_0)}$ +ε $\mathbf{N_p(\mu_1, \Sigma_0)}$

Here $\boldsymbol{\Sigma_0}$ is the homogenous covariance matrix of size pxp with the diagonal elements being 1 and the off diagonal elements are 0.9 following Alfaro and Ortega(2007).

Simulation and winsorization was done using the R package MASS for multivariate normal distribution and using the R package rrCov for computing the comedian. Simulation for the multivariate skew t distribution and g and h distributions using the R packages EMMIXskew and gk respectively. The UCL of the control chart so obtained was compared with another set of observations generated and conclusions can be drawn.

The study was to construct control charts for nonnormal multivariate data and draw conclusions. Simulation of 5000 observations was done from $\mathbf{N_p(0, I_p)}$ with p=6. Winsorization of these observations was done and bootstrapping the winsorized observations with B=1000 yielded the required 99th percentile value. Its average of the percentiles using the median as an average was obtained. This value was used as the UCL of the control chart. Any value of T^2, obtained using the robust estimators calculated from the values obtained in phase I, calculated in phase II which is greater than the UCL is said to be out of control.

Similarly, a simulation of 5000 observations each was done using multivariate skew t distribution and the g and h distribution. The procedure defined above was repeated for these observations also.

UCL values for multivariate skew t distribution for varying values of p.

Number of variables(p)	$UCL_{0.01}$	$UCL_{0.05}$
7	2.9183	1.965512
8	2.917445	1.98181
9	2.84961	1.947831
10	2.80191	1.916743
11	2.828841	1.9196
12	2.84066	1.92095
13	2.825278	1.9242
14	2.8166	1.96219
15	2.791509	1.88

UCL values for multivariate g and h distribution for different values of g and h

g(skewness)	0.5	0.75	0.99	0.5	0.75	0.99
h(elongation)	0.25	0.25	0.25	0.75	0.75	0.75
UCL	5.0065	4.960063	4.840608	5.288651	5.305734	5.189159

A comparison of these charts can be studied by varying the sample sizes and the values of the parameters of the distributions considered. The performance of these control charts compared to the traditional control charts can be investigated by varying the values of the measures of skewness and kurtosis in the case of g and h and skew t distributions.

5. Conclusion

For the skew t distribution, we observe that, as the number of variables increase, the value of the UCL reduces. In the case of the g and h distribution, for a fixed value of h, if the value of g varies, the value of the UCL reduces and when the value of h increases, again the same pattern is observed.

References

[1] Anderson T.W, (1968) An Introduction to Multivariate Statistics, New York, Wiley.

[2] Bersimis, S., S. Psarakis, and J. Panaretos. 2007. Multivariate statistical process control charts: Anoverview. Quality and Reliability Engineering International 23 (5):517–43.

[3] Chakraborti, S., P. Van der Laan, and S. T. Bakir. 2001. Nonparametric control charts: an overview and some results. Journal of Quality Technology 33 (3):304–15.

[4] Chou, Y.-M., R. L. Mason, and J. C. Young. 2001. The control chart for individual observations from a multivariate non-normal distribution. Communications in Statistics-Simulation and Computation 30 (8–9):1937–49.

[5] Devlin S.J, Gnanadesikan R and Kettenring J.R, Robust estimation of dispersion matrices and principal components, 1981, Journal of the American Statistical Association, 76:354-362.

[6] Efron, B., and R. Tibshirani. 1986. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical Science 1 (1):54–75.

[7] Falk. M, On MAD and Comedians, 1997,Ann. Inst. Statist. Math. , 1997, 615-644.

[8] Fan, S. K. S., H. K. Huang, and Y. J. Chang. 2013. Robust multivariate control chart for outlier detection using hierarchical cluster tree in SW2. Quality and Reliability Engineering International 29 (7):971–85.

[9] Hawkins, D. M. 1980. Identification of outliers. New York, NY:Chapman and Hall.

[10] Hotelling, H. 1947. Multivariate quality control. In Techniques of statistical analysis, ed. C.

[11] Jackson, J. E. 1959. Quality control methods for several related variables. Technometrics 1(4):359–77.

[12] Jackson, J. E., and G. S. Mudholkar. 1979. Control procedures for residuals associated with principal component analysis. Technometrics 21 (3): 341–9.

[13] Kokic P.N, Bell P.A, 1994, Optimal winsorizing cutoffs for a stratified finite population estimator, Journal of Official statistics, 10,419-435.

[14] Krzanowski W.J, 1975, Discrimination and classification using both binary and continuous variables, Journal of the American Statistical Association, 70, 782-790.

[15] Krzanowski W.J, 1979, Some linear transformations for mixtures of binary and continuous variables with particular reference to linear discriminant analysis, Biometrika, 66, 33-39.

[16] Lachenbruch P.A and Goldstein M, Discriminant Analysis, 1979, Biometrics, 35, 69-85.

[17] Latha V and Rajalakshmi P, Location Model for Mixed data using Winsorization with comedian, Journal of computer and mathematical sciences,Vol 9 , Issue 12, December 2018.

[18] Maronna R.A and Zamar R. H, Robust estimates of location and dispersionfor high dimensional data sets, 2002, Technometrics, 44:307-317.

[19] Martinoz F.C, Haziza D and Beaumont J.F, A method of determining the winsorization threshold, with an application to domain estimation, 2015, Survey methodology, 41-1:57-77.

[20] Montgomery, D. C. 2005. Introduction to statistical quality control. 5th ed. New York, NY:Wiley.

[21] OlkinI, Tate R.F, Multivariate correlation models with mixed, discrete and continuous variables, 1961, Annals of mathematical statistics, 32, 448-465.

[22] Pena D and Preito F.J, Multivariate outlier detection and robust covariance matrix estimation, 2001, Technometrics, 43:286-300.

[23] Phaladiganon, P., S. B. Kim, V. C. Chen, and W. Jiang. 2013. Principal component analysis-based control charts for nonnormal distributions. Expert Systems with Applications 40 (8):3044–54.

[24] Rousseeuw P and K.van Driesen, A fast algorithm for the minimum covariance determinant estimator, 1999, Technometrics, 41:212-223.

[25] Sajesh T.A and Srinivasan M.R, Outlier detection for high dimensional data using the Comedian approach, 2012, Journal of statistical computation and simulation, 82:5,745-757.