ECONOMIC DESIGN OF STATISTICAL PROCESS CONTROL USING PRINCIPAL COMPONENTS ANALYSIS AND THE SIMPLICIAL DEPTH RANK CONTROL CHART

Vadhana Jayathavaj
Rangsit University, Thailand
vadhana.j@rsu.ac.th

Adisak Pongpullponsak
King Mongkut's University of Technology Thonburi, Thailand
adisak.pon@kmutt.ac.th

Abstract:
The principal components analysis (PCA) and the simplicial depth rank control chart (r chart) have been introduced as a nonparametric multivariate statistical process control by transforming the highest principal components (PCs) with cumulative eigenvalues more than 60% to the simplicial depth rank uniform distribution. The correlated trivariate standard normal distribution with the combinations of shift in 0,1,2 and 3 times of standard deviation are simulated. The results show that the Chi-square goodness of fit test rejected the null hypothesis which the simplicial depth rank distribution is a uniform distribution. The transition matrix for variable parameters PCA r chart is simulated from the predetermined action limit (k) and warning limit (w). The dual control schemes using the reference data set (RDS) of n observations for the regular control scheme (1) and the tight control scheme (2) are RDS_1 n=150 and 300, RDS_2 n=150 and 300, number of observations $(n_1,n_2)=(1,1)$, sampling intervals $(h_1,h_2)=(1$, 19 values ranging from 0.05 to 1.0 with increment in each step 0.05), the warning region for regular scheme $(k_1-w_1)=(0.05-0.10, 0.075-0.15)$, and for the tight scheme $(k_2-w_2)=(0.05-0.10, 0.075-0.15, 0.10-0.15)$, within 22 mean shift combinations from total of 64 (4×4×4). The minimum economic cost per time unit (ECTU) using the Lorenzen and Vance's cost parameters is approximately 937.39.

Keywords: variable parameters, simplicial depth rank, quality control, principal components analysis, economic design

1. INTRODUCTION

For the process with correlated control variables, the principal components analysis (PCA) is the linear transformation from the original set of variables to the new set of variables and produces the principal components that have the most variance (the first component) to the least variance (the last component). The dimensions reduction can be done by using only some new variables from PCA process which produce the cumulative sum of variances more than 60% of total variance (Beltran, 2006). The simplicial depth is the new measure in quantifying that how well or how deep the new observations conform to the group of historical or reference data set (Liu, 1990). The simplicial depth rank is the standardized or normalized degree of the group membership and it converges to a uniform distribution in [0,1] range (Liu & Singh, 1993).

The principal components analysis (PCA) and the simplicial depth rank control chart or "PCA r chart" applied to MSPC with the small historical dataset were shown in numerical examples of Beltran's dissertation. This proposed technique created a nonparametric control chart using the simplicial depth ranks of the first and last set of principal components by means of the highest variance (lowest correlation) and the highest correlation (lowest variance) to improve signal detection in multivariate quality control with no distributional assumptions of the process control variables and the author suggested that it can be improved in cost savings and quality improvement (Beltran, 2006).

In extending the research on PCA r chart, the purposed studies are as follows.

- Verify the uniform distribution of the simplicial depth rank statistic.
- Evaluate the average run length (ARL) of PCA r chart.
- Develop the economic design for variable parameters PCA r chart.

2. LITERATURE REVIEW

2.1. The principal components analysis simplicial depth rank control chart

Principal component analysis has been used as MSPC tool for detecting faults in process with highly correlated variables. Principal components obtained from covariance and correlation matrices are different. Variables should be standardized if they are measured on scales with widely differing ranges or if the unit of measurements are not commensurable (Johnson & Wichern, 2002). A data depth measures how deep (or central) a given point relative to a given data cloud. Serfling (2004) reviewed depth functions in nonparametric multivariate inference and mentioned that the simplicial depth (Liu, 1988; Liu , 1992) opened up the potential of "depth functions" as a new methodology both powerful and broad in nonparametric inference.

Beltran (2006) proposed the new multivariate nonparametric control chart using the principal components analysis (PCA) in dimensional reduction, only the first principal components ((cumulative variance from the highest variance) and the last principal components (highest correlation) are computed the simplicial depth rank and plotted on the r chart with Type I error = 0.05 (Liu, 1995).

Using independent univariate chart are not always the best method for monitoring correlated characteristics, because the correlations between variables result in degrading the statistical performance of these charts. The false alarm rate or probability of Type I error is increased if each variable is controlled separately. (Asem Khalidi, 2007).

The PCA r chart was applied to 5 normally distributed variables and provided the numerical simulation to compare the control result with Hotelling T^2 and concluded that the new technique was more favorable ARL performance than Hotelling T^2 for the small (+1σ), medium (+2σ) , and large (+3σ) shift in the mean of the first original variable (Zarate & Okogbaa, 2007).

The further study in economic design of PCA r chart will be confirmed their effectiveness.

2.2. The simplicial depth

Let 1 denote the set of n observations in p-dimensional space. Liu (1995) proposed the nonparametric control charting scheme which is based on ranking data depth of the multivariate observations of the p

process variables and plotting the ranks using the univariate control chart in order to easily detect multivariate process shift visually. Liu (1990) defined simplicial depth relative to the probability that a point lies within a random simplex with vertices $p+1$. These $p+1$ vertices are independent observations from the distribution say F(not the F-distribution). $D_F(x)$ or $D(x)$ represents the measure of depth of a point x with respect to the continuous distribution F. For some given point x in p-dimensional space $\mathbf{R}^p$, the simplicial depth is a measure of how central x lies within a random sample $\{X_1, X_2, \dots, X_n\} \subset \mathbf{R}^p$.

The simplicial depth of x with respect to a continuous distribution F as $D(x) = D_F(x) = P_F\left(x \in s\left[X_1, X_2, \dots, X_{p+1}\right]\right)$ where $s\left[X_1, X_2, \dots, X_{p+1}\right]$ represents a $p+1$ dimensional simplex with vertices $X_1, X_2, \dots, X_{p+1}$ which are random observations from F. The s for univariate, bivariate and trivariate cases represent the line segment, the triangle and the tetrahedron, respectively. This measure $D(x)$ describes "how central the point x is within the distribution".
When the distribution is unknown, the reference sample $X_1, X_2, \dots, X_n$ is used to compute the simplicial depth. The sample simplicial depth is

$$D_n(x) = \binom{n}{p+1}^{-1} \sum_{1 \le i < p+1 \le n} I\left(x \in s\left[X_1, X_2, \dots, X_{p+1}\right]\right) \tag{1}$$

in which F_m represents the empirical distribution of $X_1, X_2, \dots, X_n$ with $n \ge (p+1)$ and $I(.)$ is the indicator function

$$I\left(x \in s\left[X_1, X_2, \dots, X_{p+1}\right]\right) = \begin{cases} 1 & \text{if } x \in s\left[X_1, X_2, \dots, X_{p+1}\right] \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

2.3. Liu's r chart

Let G denote the prescribed quality distribution with p process variables with $Y_1, Y_2, \dots, Y_n$ random observations which represent the reference in control data set. Collect a sample of new observations $X_1, X_2, \dots, X_t$ and assume that its distribution is F. If the process has gone out of control, compare the sample of new observations $X_1, X_2, \dots, X_t \sim F$ against the reference sample $Y_1, Y_2, \dots, Y_n \sim G$.

$$H_0: G = F \text{ with a false alarm rate of } \propto$$

$$H_a: G \ne F \text{ There is a shift in location } \frac{\text{and}}{\text{or}} \text{ or scale increase from } G \text{ to } F.$$

Compute the ranks of the depths $Y \sim G$. As the rank decreases, the more outlying is the point within that distribution.

$$r_G(y) = P\{D_G(Y) \le D_G(Y) | Y \sim G\} \tag{3}$$

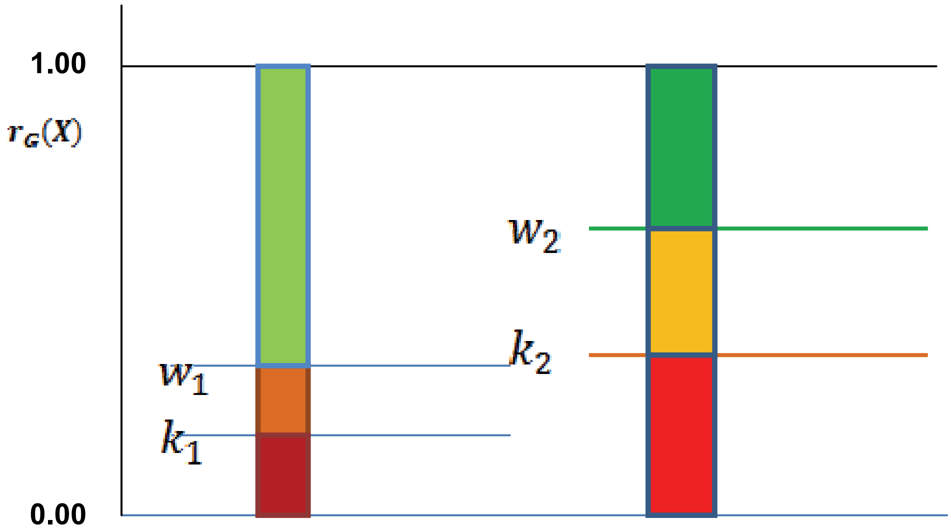
$$r_{G_n}(y) = \frac{\#\{Y_j | D_{G_n}(Y_j) \le D_{G_n}(y)\}}{n} \quad \text{for } j = 1, 2, \dots, n \tag{4}$$

The assumption is that the distribution $D_G(X)$ is continuous and that $r_G(X)$ converges to a uniform distribution $\sim U[0,1]$), as such the expected value 0.5 which will serve as the center line (CL). The process will be declared out of control, when $r_G(X_i) < \propto$ (Liu, 1995).

2.4. Variable parameters control charts

The variable parameters (Vp) control chart is aiming to detect assignable causes effectively, the process has to be monitored between predetermined sampling schemes depending on the sample plot is on what area of the control chart. The sampling scheme of the Vp $\bar{X}$chart is to use the small sample (n_1) size , the long sampling interval (h_1) and the wide action limit coefficient (k_1) as long as the sample point is close to the target so that there is no indication of process change. However, if the sample point is close to, but still within the action limits so that there is some indication of process shift, then the large sample size (n_2), the short sampling interval (h_2) and the narrow action limit coefficient k_2 are used (De Magalhaes, Epprech & Costa, 2001). In the PCA r chart, the simplicial depth rank statistic ($r_G(X)$) is a uniform distribution in the 0 to 1 range, the control regions of variable parameters control chart can be designed as shown in Figure 1.

Figure 1: The variable parameters PCA r chart control schemes



3. METHODOLOGY

In order to complete the objectives for this study, the computation flow chart is shown in Figure 2. An important input for the expected cost per time unit (ECTU) calculation is their transition matrix (De Magalhaes, Epprecht & Costa, 2001, pp. 191-200).

The Transition Matrix for Economic Model

Given that the subscript 1 is for the regular control scheme and 2 for the tight control scheme

p_{11} = probability that the simplicial depth rank r is place in the central region of the regular control scheme $(r > k_1)$, the next observation will be tested hypothesis against the regular control scheme

p_{12} = probability that the simplicial depth rank r is place in the warning region of the regular control scheme $(w_1 < r \le k_1)$, the next observation will be tested hypothesis against the tight control scheme

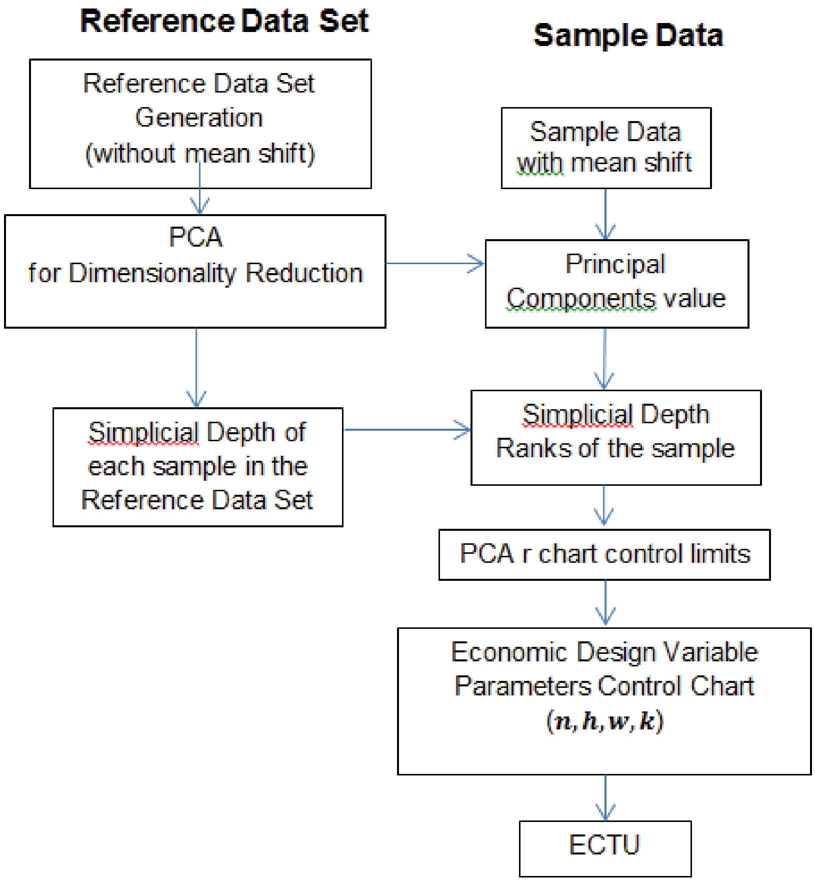
p_{21} = probability that the simplicial depth rank r is place in the central region of the tight control scheme $(r > k_2)$, the next observation will be tested hypothesis against the regular control scheme

p_{22} = probability that the simplicial depth rank r is place in the central region of the regular control scheme $(w_2 < r \le k_2)$, the next observation will be tested hypothesis against the tight control scheme

Find the Transition matrix $\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix}$

- Generate random number of $\boldsymbol{x}$ as per its distribution with mean, standard deviation and correlation structure for M observations
- Find the simplicial depth rank $r_G(v)$ or *r* with respect to the given reference data set.
- Tally r values into the given interval of *r* and given F.
- Divide F by M given p, this p is the probability of that given *r* interval.

Figure 2: Economic design of variable parameters PCA r chart



Economic Cost per Time Unit (ECTU)

The economic cost per time unit (ECTU) follows the Lorenzen & Vance Model (Lorenzen & Vance, 1986).

The process variables can be shifted in δ times of standard deviation. The level of shifts (δ_j) of the three variables (j=1,2,3) are 0,1,2, and 3 times of standard deviation. Compute the ECTU for each shifts subject to RDS n, δ, k, w, n, h of the regular and tight control schemes. All the computations and figures are run by MATLAB, 7.6.0(R2009a) and MINITAB 16.

4. NUMERICAL RESULT

4.1. Input data generation

The process data have 3 variables (x_1,x_2,x_3) with standard normal distribution. Their correlations are predetermined with high correlation for (x_1,x_3) at 0.80 and low correlation for (x_1,x_2) at 0.30. Once x_1 is correlated with both x_2 and x_3, it is inevitably to have some correlation between x_2,x_3 , and it is given at 0.24.

4.2. Principal components analysis

Random numbers with sample size 50, 100, 150, 200, and 300 are generated using the trivariate standard normal distribution with their correlations from 4.1, the eigenvectors and cumulative eigenvalues from PCA are shown in Table 1. The first principal component with variance more than 60% come from the sample size of 100 up.

Table 1: Principal Component Analysis with eigenvectors and cumulative eigenvalues.

	PCA	Sample size 50	100	150	200	300
v_1	x_1	0.7386	0.6911	0.6541	0.6683	0.6668
	x_2	-0.0988	0.1135	0.4502	0.3265	0.3993
	x_3	0.6669	0.7138	0.6078	0.6684	0.6292
v_2	x_1	0.1188	-0.0595	-0.1994	-0.1220	-0.2870
	x_2	0.9928	0.9932	0.8778	0.9345	0.9168
	x_3	0.0154	-0.1003	-0.4355	-0.3344	-0.2777
v_3	x_1	-0.6636	0.7204	0.7296	0.7338	-0.6878
	x_2	0.0678	-0.0268	-0.1636	-0.1419	0.0046
	x_3	0.7450	-0.6931	-0.6640	-0.6644	0.7259
	λ_1	58.37	66.58	63.18	66.82	63.31
	$\lambda_1+\lambda_2$	87.93	92.85	92.98	93.93	93.28
	$\lambda_1+\lambda_2+\lambda_3$	100.00	100.00	100.00	100.00	100.00

4.3. The simplicial depth rank probability distribution

The reference data sets (RDS) n=50,100,150,200,300, and 500, the simplicial depth rank for 100000 random numbers are generated, their probability distribution by 10 intervals of 0.1 each are shown in Table 2. All of the simulated probability distributions of the simplicial depth rank are rejected that the simplicial depth rank distribution is uniform by the Chi-Square goodness of fit test, the p value = 0.0000 for every RDS n.

Table 2: Probability distribution of the rank of the simplicial depth rank ($r_{G(X)}$) from 100000 random numbers simulation.

n			50	100	150	200	300	500
from		to						
0.9	-	1.0	0.1459	0.2186	0.1603	0.1236	0.1647	0.1100
0.8	-	0.9	0.0844	0.0984	0.0861	0.1069	0.0801	0.0963
0.7	-	0.8	0.0322	0.0923	0.1336	0.0746	0.1150	0.1105
0.6	-	0.7	0.1618	0.0812	0.1214	0.0640	0.1082	0.0925
0.5	-	0.6	0.0758	0.1194	0.0848	0.0813	0.1105	0.0937
0.4	-	0.5	0.1299	0.0748	0.0583	0.1262	0.0893	0.1024
0.3	-	0.4	0.0412	0.1427	0.1143	0.0936	0.0917	0.1055
0.2	-	0.3	0.2136	0.0793	0.0780	0.1164	0.0850	0.1004
0.1	-	0.2	0.0883	0.0611	0.0606	0.1552	0.0971	0.1013
0.0	-	0.1	0.0268	0.0323	0.1026	0.0582	0.0585	0.0874
		Total	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

4.4. Possible mean shift settings

Each process variable mean can be shifted in δ times of standard deviation. The level of shifts (δ_j) of the three variables (j=1,2,3) are 0,1,2, and 3 times of standard deviation. The total combination of mean shifts are 4x4x4 = 64 shifts. The 22 mean shifts are selected in computations: no variable mean shift (0,0,0), one variable mean shift (δ_i), two variables mean shift (δ_i, δ_j), and all three variables mean shift (δ_i, δ_j, δ_k) with the number of combination of mean shifts are 1, 9, 9, and 3 respectively as shown in the first four columns of Table 3.

4.5. PCA r chart performance

The simplicial depth rank probability from 100000 simulated observations and direct average run length (ARL) from 30000 simulated run lengths using reference data set *n*=150 and 300 and 22 selected mean shifts against the action limit k=0.05 are simulated. The ARLs are computed from the Type I error (α) and Type II error (β) probabilities from simulation and compare with the ARL from direct simulation in each mean shift. Overall ARL performances are the larger mean shift, the lesser the ARL, and at some mean shift: ARL from probability simulation are quite large different from ARL from direct simulation.

4.6. Variable parameters economic design

Cost Parameters

From the numerical example of Lorenzen & Vance (1986) in ECTU computations, the model input parameters were the example from foundry operations, where periodic samples of molten iron are taken to monitor the carbon-silicate content of the casting. The values of the input variables are $G = T_* = \frac{5}{60}$ hours; $T_2 = \frac{45}{60}$ hours; $1/\lambda = 50$; $C_0 = \$114.24$/hour; $C_1 = \$949.20$/hour; $Y = W = \$977.40$; $a = 0$; $b = 0$; $\delta_1 = 1$; $\delta_2 = 0$. n_1 =, n_2 =1, h_1 and h_2 vary from 0.05 to 1.00 = 19 values (0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85,0.90, and 1.00) (De Magalhaes, Epprecht, Costa, 2001, pp. 191-200).

Variable Parameters Control Schemes

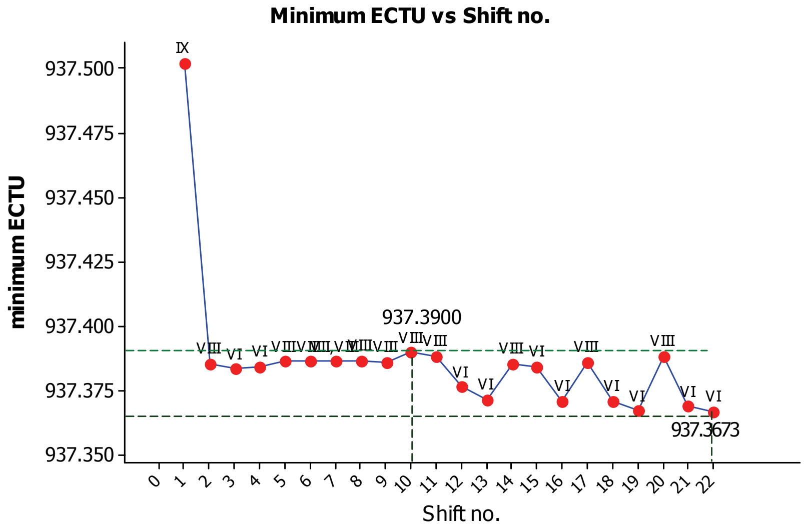
The combination of reference data sets and control limits of 9 dual variable parameters schemes for the simplicial depth rank r control chart are shown in Table 3.

Table 3: The PCA r chart control schemes

Region	Regular [R]	Tight 1 [T1]	Tight 2 [T2]
Central	0.10-1.00	0.15-1.00	0.15-1.00
Warning	0.05-0.10	0.075-0.15	0.10-0.15

Regular	Tight	Vp PCA R chart
150R	150T1	I
150R	150T2	II
150R	300R	III
150R	300T1	IV
150R	300T2	V
150T1	150T2	VI
300R	300T1	VII
300R	300T2	VIII
300T1	300T2	IX

Figure 3: Minimum ECTU of 22 mean shifts of Variable Parameters PCA r chart



4.7. Minimum Expected Cost per Time Unit (ECTU)

The minimum Expected Cost per Time Unit (ECTU) of 22 mean shifts from the 9 dual scheme variable parameters PCA r chart are shown in Figure 3.

5. CONCLUSION

The results from an economic design of the variable parameters PCA r chart using the first components with cumulative variance more than 60% of this correlated trivariate case are as follows, The simulation of 100000 random numbers, the simplicial depth rank statistic ($r_{G(X)}$) is not conformed to the uniform distribution as proved by Liu & Singh (1993) that the simplicial depth rank converges to the uniform distribution.

The average run length (ARL) of PCA r chart from simulations are also different from Type I error probability (α) which simulated from the given distribution and correlations (this case α is not assumed to be a uniform distrbution).

For the PCA r chart, the transition matrix for economic design comes from simulation. The ECTU of RDS n=150 and RDS n=300 for RDS$_1$ n and RDS$_2$ n have no addditional costs because they are selected from process historical data. The simplicail depth rank uses only one observation in each sampling (n_1,n_2)=(1,1) . The variable parameters combinations shows the ECTU is approximately at 937. This ECTU is in the same range as in the previous Shewhart $\bar{X}$ chart study (Pongpullponsak, Suracherdkiatchai & Panthong, 2009).

REFERENCE LIST

1. Asem Khalidi, M. S. (2007). Multivariate Quality Control: Statistical Performance and Economic Feasibility . *Ph.D. Dissertation*. Wichita State University.
2. Beltran, L. A. (2006). Nonparametric multivariate statistical process control using principal component analysis and simplicial depth. *Ph.d. DISSERTATION, College of Engineering and Computer Science, The Unversity of Central Florida*. Orlando, Florida.
3. De Magalhaes, M. S., Epprecht, E. K. & Costa A. F. (2001). Economic design of a Vp QUOTE chart. *International Journal of Production Economics*, 74(2001),191-200.
4. Johnson, R. A., & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis.* Upper Saddle River, NJ, USA: Pearson Education International.
5. Liu, R. Y. (1988). On a notion of simplicial depth. *Proceedings of the National Academy of Science USA*, *85*, 1732-1734.
6. Liu, R. Y. (1990). On a notion of data depth based on random simplices. *The Annals of Statistics, 18*, 405-414.

7. Liu, R. Y. (1995). Control charts for multivariate processes. *Journal of the American Statistical Association, 90*(432), 1380-1387.
8. Liu, R. Y., & Singh, K. (1993). A Quality Index Based on Data Depth and Multivariate Rank Tests. *Journal of the American Statistical Association, 88*, 257-260.
9. Lorenzen, T. J., & Vance, L. C. (1986). The economic design of control charts: A unified approach. *Technometrics*, 28:3-10.
10. MINITAB Thailand. (2010, February 2). MINITAB 16 Order Number 100004968850, Single License.
11. Pongpullponsak, A., Suracherdkiatchai, W., & Panthong, C. (2009). The economic model of X bar control chart using Shewhart method for skewed distributuions. Thailand Statistician *Journal of the Thai Statistician Association*, 7(1), 81-99.
12. Serfling, R. (2004). Depth functions in nonparametric multivariate inference. *DIMACS Series in Discrete Mathematics and Theoritical Computer Series*, 1-16.
13. The Math Works. (2009, February 12). MATLAB 7.6.0 (R2009a). License Number 350306.
14. Zarate, P. B., & Okogbaa , O. G. (2007-1014). *A Non-Parametric Control Chart for Multivariate Processes Using Principal Component Analysis*. CiteSeerX,The Pennsylvania State University. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/versions?doi=10.1.1.19.6967