العنوان:	Multivariate Statistical Process Control Based on Principal Component Analysis : MSPC-PCA
المصدر:	المجلة المصرية للدراسات التجارية
الناشر:	جامعة المنصورة - كلية التجارة
المؤلف الرئيسي:	S., Hanaa M.
مؤلفين آخرين:	B Ashraf A.، El Bayomy, A. T.(Auth.)
المجلد/العدد:	مج37, ع1
محكمة:	نعم
التاريخ الميلادي:	2013
الصفحات:	57 - 82
رقم MD:	660235
نوع المحتوى:	بحوث ومقالات
قواعد المعلومات:	EcoLink
مواضيع:	العمليات الإحصائية، التحليل الإحصائي، إدارة التحكم
رابط:	http://search.mandumah.com/Record/660235

Multivariate statistical process control Based on principal component Analysis (MSPC-PCA)

Hanaa M .S, EL- Bayomy A. T and Ashraf A. B

Faculty of Commerce, Mansura University

Abstract

Tracking batch to batch variation and detecting abnormal events at early stages of a batch run is of critical importance in chemical process and many other industries which employ batch-wise operations. Multivariate quality control chart (MQC) Hotelling's T^2 ,the multivariate exponentially weighted moving average (MEWMA) chart, multivariate cumulative sum (MCUSUM) chart, and the multivariate process variability control chart have been a key technique for this purpose. A significant step forward in recent years in multivariate statistical process control (MSPC) for operational condition monitoring and fault diagnosis has been the introduction of principal component analysis (PCA) for compression of process data / significant

progress has been made since Kresta ,et al introduced the idea of using principal component analysis (PCA) to pre - processing the process variables. PCA provides a smaller number of latent variables that can be used to replace the original observed variables in calculating the T^2 and other charts. Each principal component (PC) is the linear combination of the original variables with the first PC captures the majority of variation in the data, and the second PC captures the majority of remaining variation and is orthogonal to the first PC, and so forth. There has been a wealth of publications in this area over the last ten years, and the work has been reviewed by a number of researchers.

Key Words: Multivariate statistical process control Based on principal component Analysis (MSPC-PCA)

l. Introduction

In modem manufacturing processes , massive amounts of multivariate are routinely collected through automated in-process sensing. These date often exhibit high correlation rank deficiency , low signal to- noise ratio and missing values , A significant step for word in recent years in multivariate statistical process control (MSPC) for operational condition monitoring and fault diagnosis has been the introduction of principal component analysis (PCA) for compression of process date .This chapter discusses these issues and indicates the use of multivariate statistical process control based on principal component analysis (MSPC - PCA) as an efficient statistical tool for process understanding (PCA) (Jackson , 2003) are used to reduce the dimensionality of the monitoring space by projecting the information in the original variables down onto low-dimensional subspace defined by a few latent variables the process is then monitored in these latent subspace by using a few multivariate control charts built from multivariate statistics which can be thought of as process performance indices, or process wellness indices.

2. (MSPC - PCA)

The MSPC - PCA monitoring scheme , as any SPC scheme is carried out in two phases (model building) monitoring charts are built according to a set of historical in - control date, once the

performance of the process has been understood and modeled, stability are assumption of its behavior and process stability are checked . In phase II (model exploitation) these charts are used to monitor the process using on-line date, assuming the form of the distribution to be known along with its values of the in-control parameters (Woodall, 2000).

2.1 phase 1 (Model building):

Exploratory Date Analysis and off - line process monitoring

The main goal in phase I is to model the in-control process performance based on a set of historical in control (reference) date. This date set (HDS) is one in which the process has been operating consistently (stable over time) in an acceptable manner , and in which only good quality products have been obtained occasionally this historical in-control date set is not directly available but has to be extracted from historical databases in an iterative fashion as commented bellow. This explorative analysis of historical database is a useful technique for improving process understanding and detecting past faults in the process (out -of -control samples). By correctly diagnosing their root causes , some countermeasures can be implemented, optimizing the future performance of the process Consider that historical database consists of a set of N multivariate observation (objects or samples) on K variables (on-line process

measurement, dimensional variable or product quality date) arrangement in a (N x K) date matrix Z variable in matrix Z are often pre-processed by mean -centering and scaling the average value of each variable is calculated and then subtracted from the date. This usually improves the interpretability of the model because all pre-processed variables will have mean value zero . By scaling to unit variance each original variable is divided by its standard deviation and will have unit variance . Given that projection methods are sensitive to scaling ,this is particularly useful when the variables are measured in different units . After preprocessing, matrix Z is transformed into matrix X. Principal Component Analysis (PCA) is used to reduce the dimensionality of the process by compressing the high- dimensionality original date matrix X into a low-dimensional subspace of dimension A ($A \leq$ rank (X)), in which most of the date variability is explained by a fewer number of latent variables, which are orthogonal and linear combinations of the original ones . This is done by decomposing X into a set of A rank 1 matrices

$$X = \sum_{a=1}^{A} t_a P_a^T + \sum_{a=A+1}^{rank(X)} t_a P_a^T = TP^T + E = \hat{X} + E \rightarrow (1)$$

P (K x A) is the loading matrix containing the Loading vectors P_a, which are the eigenvectors , corresponding to the A largest eigenvalues of the covariance matrix of the original pre-treated date set X , and define the directions of highest variability of the new latent A-dimensional subspace T (NxA) is the score

matrix containing the location of the orthogonal projection of the original observations onto the latent subspace. The columns t_a of the score matrix T

$$XP_a = ta \qquad (2)$$

represent the new latent variables with variances given by their respective eigenvalues λ_a. The λ_a are a measure of the amount of variance described by the t_a ,p_a pair .In the context, we can think of variance as information .Because the t_a ,p_a pairs are in descending order of λ_a ,the first pair capture the largest amount of information of any pair of the decomposition . In fact it can be shown that t_1 ,p_1 pair capture the greatest amount of variation in the data that it is possible to capture with a linear Component .Subsequent pairs capture the greatest possible variance remaining at that step .

To concept of principal component is shown graphically in figure (1) .The figure shows a three dimensional data set where the data lie primarily in a plane , thus the data is well described by a two principal component (PC) model .The first eigenvectors of PC aligns with the greatest variation in the data while the second PC aligns with the greatest amount of variation that is orthogonal to the first PC.

Generally it is found that the data can be adequately described using far fewer Principal Components than original variables .

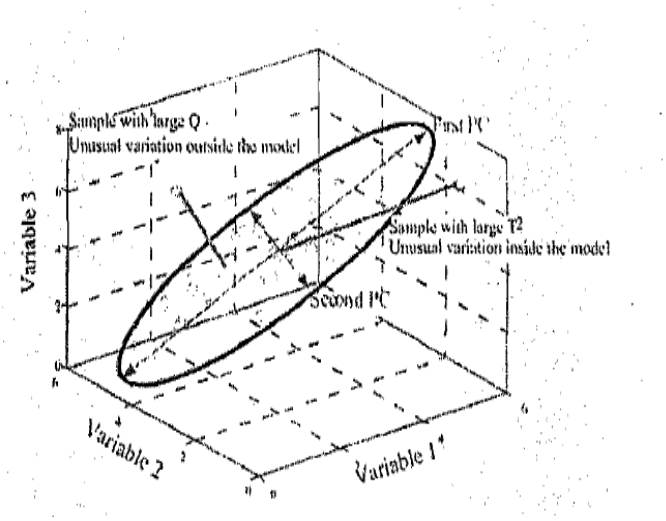

Fig.1. Principal Component Model of Three Dimensional Data Set Lying Primarily in a Single Plane Showing Q and T^2 Outliers.

The new latent variable summarize the most important information of the original K variables , and thus can predict (reconstruct) X with minimum mean square error $X = TP^T$ .Matrix E (NxK) contains the residuals (statistical noise) such that the information that east not explained by the PCA model The dimension of the latent variable subspace is often quite small compared with the dimension of the original variable space .

Equation (2) shows that the PCA model transforms each K- dimensional original observation vector X_i (ith row of matrix X) into an A-dimensional score vector $t^Ti = \{t_{i1}, t_{i2},..., t_{iA}\}$

(ith row of matrix T). From the scores and the residuals (predication errors) associated with each observation, two complementary (orthogonal or independent) statistics are derived: the Hotelling's T^2_A and the SPE (sum of squared prediction errors), the T^2_A statistic for the ith observation is defined as :

$$T_A^2 = t_i^T \Theta^{-1} t_i = \sum_{a=1}^{A} \frac{t_a^2}{\lambda_a} \qquad \rightarrow (3)$$

Where Θ (A X A) is the covariance matrix of T (diagonal matrix of the highest A eigenvalues { λ_1, ……, λ_A}). This statistic is the Hoteling's T^2 statistic when a reduced subspace to the projection of an operation onto this subspace under the assumption that the score follow a multivariate normal distribution (they are linear combination of random variables) it holds (Tracy et al , 1992) that in phase I , T^2 (times a constant) has a beta (B) distribution

$$T_A^2 \sim \frac{(N-1)^2}{N} B_{A/,(N-A-1)/2} \qquad \rightarrow \qquad (4)$$

while in phase II , T^2_A (time a constant) follows an F distribution

.

$$T_A^2 \sim \frac{A(N^2-1)}{N(N-A)} F_{A,(N-A)} \quad \rightarrow \quad (5)$$

The difference in both distribution comes from the fact that in phase I, the same observation vectors X_i collected in the reference data set are used for two purposes :

(1) to build the PCA model and work out the control limit of the charts.

(2) to check whether they fall within these control limits . Therefore, observation in the reference data set are not independent of PCA model parameters used to derive the statistics to be monitored . In contrast , in phase II new observation (not used for model building) are checked against the control limits calculated from the in-control data , and therefore , independence is guaranteed. Anyway, if a large reference data set is available Eq. (5) can also be used for approximating the distribution of the T^2_A statistic in phase I.

On the other hand the SPE statistic for ith observation X_i , is given by:

$$SPE=Q_i = e_i e_i^T = \left(X_i - \hat{X}\right)(X_i - \hat{X})` \quad \rightarrow (6)$$

Where e_i , is the residual vector of ith observation, and X_i is the predication of the observation vector X_i , from the PCA model . The SPE statistic represents the squared Euclidean

(perpendicular) distance of an observation from this subspace , and gives a measure of how close the observation is from the A-dimensional subspace. Assuming that residuals follow a multivariate normal distribution Jackson and . Mudholkar (1979) and Eriksson et al (2001) derived approximate distributions for such quadratic forms.

From these two statistics, in MSPC-PCA complementary multivariate control charts are constructed .Shwhart- type control charts for individual observation are often used in practice . Nevertheless, other types of rational subgrouping or multivariate charts such as Multivariate EWMA (MEWMA) charts (Lowy et al . 1992) may be used. The latter may be specially suited for autocorrelated processes.

The control limits of the multivariate control charts are calculated following the traditional SPC philosophy . in phase I , an appropriate historical or reference set of , data (collected from one or various periods of plant operation when performance was good) is chosen which defines the normal or in- control operation conditions for a particular process corresponding to common-cause variation. The in-control PCA model is then built on these data . Any periods containing variations arising from special events that one would like to detected in the future are omitted at this stage . The choice of the reference (in-control) data set is critical to the successful application of the procedure (Kourti and

Macgregor 1996). Control limits for good operation in the control charts are defined based on this references data set. In phase II, values of future measurement are compared against these limits .

Upper control limited (UCL) for the Shewhart T^2_A chart at significance level (type I risk) a can be obtained for phase I from Eq. (4)

$$UCL\,(T_A^2)_\alpha = \frac{(N-1)^2}{N} B_{(A/2,(N-A-1)/2),\alpha} \quad \rightarrow \quad (7)$$

Where $B_{(A/2,\ (N-A-1)/2)\alpha}$ the 100 (1 - α)% percentile of the corresponding beta destruction that can be computed from the 100(1 -α) % percentile of the corresponding F distribution by using the relationship (Tracy et., al 1992)

$$B_{(A/2,(N-A-1)/2),\alpha} = \frac{(A/(N-A-1))\,F(A,N-A-1)_\alpha}{(1+(A/(N-A-1))\,F(A,N-A-1)_\alpha} \quad \rightarrow (8)$$

For phase II, the corresponding UCL, from Eq . (5) is given by :UCL

$$(T_A^2)_\alpha = \frac{A(N-1)^2}{N(N-A)} F_{(A,(N-A)),\alpha} \quad \rightarrow \quad (9)$$

Regarding the UCL for the Shewhart SPE chart several procedures can be used , Jackson and Mudholkar (1979) showed that an approximate SPE critical value at significance level a is given by :

$$UCL(SPE)_{\alpha} = \theta_1 \left[\frac{Z_{\alpha}\sqrt{2\theta_2 b_0^2}}{\theta_1} + 1 + \frac{\theta_2 b_0 (b_0 - 1)}{\theta_1^2} \right]^{1/b_0} \quad \rightarrow \quad (10)$$

Where $\theta_K = \sum_{j=a+1}^{rank(x)} (\lambda_j)^k, b_0 = 1 - 2\theta_1\theta_3/3\theta_2^2$, λ_j are the

eigenvalues of the PCA residual covariance matrix E^TE /(N - 1), and Z_a is the 100 (1 -α) % standardized normal percentile.

Nomikos and McGregor (1995) suggested a simple and fast way to estimate the parameters g and b which is based on matching moments between a $g\chi^2_b$ distribution and the sample distribution of SPE . The mean (μ-gb) and variance ($\sigma^2 = 2g^2b$) of the $g\chi^2_b$ distribution are equated with the sample mean (b) and variance (v) of the SPE sample .Hence the upper SPE control limit at significance level a is given by :

$$UCL(SPE)_{\alpha} = \frac{v}{2b} \chi^2_{(2b^2 \ / \ v),\alpha} \quad \rightarrow \quad (11)$$

Where χ^2 $(2b^2/v),\alpha$ is the 100 (1 - α) % percentile of the corresponding chi-squared distribution.

Another ether method based on the statistical test of equality of variances from normal distribution is proposed by Eriksson et. al (2001) based on the SPE , they define the absolute

distance to the model (DModx) of an observation as its (corrected) residual standard deviation :

$$DModX = C\sqrt{\frac{SPE}{(K-A)}} \qquad \rightarrow \quad (12)$$

Where C is a correction factor (function of the number of components) to be used in phase I. This correction factor takes into account that the distance to the model (DModX) is expected to be set slightly smaller for an observation in the reference set because it has influenced the model.

This correction only matters if the number of observation in the reference set is small.

In phase II, c =1

They also define the normalized distance to the model

($DModx_{norm}$) as:

$$DMOdX_{norm} = \frac{DMOdX}{S_0} \qquad \rightarrow \quad (13)$$

Where

$$S_0 = \sqrt{\sum_{i=1}^{N}\sum_{k=1}^{K} e_{ik}^2/(N-A-1)(K-A)}$$

is the pooled residual standard deviation. This is an estimation of the residual variability taking into account all the observation used to build the model (reference data set).

Assuming that the statistic $(DModX_{norm})^2$ has an approximate F distribution with (K-A) and (N-A-1) (K-A) degrees of freedom for the in-control observations, the UCL for the Shewhert SPE chart at significance level a is expressed as :

$$UCL\ (SPE)_{\alpha} = \frac{K-A}{C^2}\ S_0^2\ F_{(K-A\ ,(N-A-1)\ (K-A)\ ),\alpha} \quad \rightarrow (14)$$

Where $F(k\text{-}A,(N\text{-}A\text{-}1)\ (K\text{-}A))_{,\alpha}$ is the 1(1- α) % percentile of the corresponding F distribution .

The normally assumption on which these calculations are based is usually quite reasonable in practice. Anyway ,control limits for multivariate charts can be obtained from distribution free methods by repeated sampling . The only requirement is to have a large in-control data set from which the external reference distribution (Box ethnic al. 1978) for any statistic can be obtained.

The two multivariate control charts (T^2_A and SPE) differ in their conceptual meaning .They are two complementary indices that provide a picture of the wellness of the process at a glance (Kourti 2005). The T^2_A chart checks if an observation project on the hyperplane defined by the latent subspace is within the limits

determined by the reference (in-control) data. Thus, a value of this statistics exceeding the control limits indicate that the corresponding observation presents abnormal extreme values in some (or all) of its original K variable , even though it maintains the correlation structure between the variables in the model. This observation can be tagged as an abnormal outlier inside the PCA model (an extremist or severe outliers) (Martens and Naes, 1989). On the other hand, the SPE chart checks if the distance (noise variation) of an observation to the latent hyperplane is inside the control limits . The SPE chart values exceeding the control limits are related to observation that do not behave in the same way as the ones used to create the model (in-control data), in the sense that there is a breakage of the correlation structure of the model . This chart will detect the occurrence of any new event that cause the process to move away from the hyperplane defined by references model. This kind of observations can be tagged as outliers outside the model (an alien or moderate as outliers) (Martens and Naes , 1989)

Severe outliers are influential observations with high leverage on the model. i.e., with strong power to pull the Principal directions toward themselves, creating fictitious Component and misleading the PCA model (Eriksson et al., 2001). Severe outliers mislead the PCA model due to the great influence that they exert during model building . Therefore model validation is extremely needed in the phase 1 stage in order to

remove from the data matrix these dangerous outliers (out-of-control) observations and afterwards, recalculate the PCA model. Anyway, before removing any observation from the data matrix, some diagnostics using contribution plots (discussed below) and process insight should be used in order to sort out false alarm outliers from the real once . This process of model building and validation is done iteratively until no multivariate control charts signals any real outlier . As a side -effect from this debugging procedure , the root causes of the out-of - control observation can be discovered importance process knowledge and future process performance.

2.2 phase II (model exploitation) on -line process monitoring

Once the references PCA model and the control limits for the multivariate control charts are obtained new process observation can be monitored on line when a new observation vectors is available after pre-processing it is projected onto the PCA model yielding the scores and the residuals , from which the value of the Hotelling's T^2_A and the value of the SPE are calculated . This way, the information contained in the original K variable is summarized in these two indices that are plotted in the corresponding multivariate T^2_A and SPE control charts. No matter what the number of the original variable K is only two points have to be plotted on the charts and checked against the control limits . The SPE chart should be checked first . If the

points remain below the control limits in both charts the process is considered to be in-control .If a point is detected to be beyond the limits of one of the charts, then a diagnostic approach to isolate the original variables responsible for the out-of-control signal is needed .In MSPC -PCA once of the most widely used approaches is the contributions plots approach (Kourti and MacGregor, 1996).

Contribution plots can be derived for abnormal points in potholes charts. If the SPE chart signals a new out-of-control observation , the contribution of each original kth variable to the SPE at this new abnormal observation is given by its corresponding squared residual:

$$Cont(SPE\,; x_{new,k}) = e^2_{new,k} = \left(x_{new,k} - \hat{x}_{new,k}\right)^2 \quad \rightarrow (15)$$

In case of using the distance to the model (DMODX) statistic , the contribution of each original kth variable to the-DMODX is given by (Eriksson et al., 2001)

$$Cont(DMODX_i\ \chi_{new,k}) = w_k\ e_{new,k} \quad .(16)$$

where w_k is the square root of the explained sum of squares for the Kth variable. Variable with high contributions in this plot should be investigated.

If the abnormal observation is detect by the T^2_A chart the diagnosis procedure is carried out in two steps (i) a bar plot of the normalized scores for that observation ($t_{new,a}/\lambda_a$) is plotted and ath score with the highest normalized value is selected ;(ii) the contribution of each original Kth variable to this ath score at this new abnormal observation is given by:

$$Cont(t_{new.a}\,; x_{new.k}) = p_{ak} x_{new,k} \quad \rightarrow \quad (17)$$

where pak is the loading of the Kth variable at component a .A plot of these contributions is created variables on this plot with high contributions but with the same sign as the score should be investigated (contributions of the opposite sign, will only make the score smaller). When there are some scores with high normalized values, an overall average contribution per variable can be calculated , over all the selected scores (Kourti 2005) .

Contribution plots are a powerful tool for fault diagnosis . They provide a list of the process variables that contribute numerically to the out-of-control condition (they are no longer consistent with normal operating conditions), but they do not reveal the actual cause of the fault . Those variables and any variables highly correlated with them should be investigated .

Incorporation of technical process knowledge is crucial to the problem and discover the root causes of the fault.

3. Aplication Study:

For the application of the multivariate quality control chart, data originate from urea production process, which consists of the three stages and the analysis of laboratory. The number of the sample is 3000 observations taken per hour. The advantages of this sample that, it has several variables and several stage of the production. This advantage of the production is the basic reason for choosing this production to allow us to study the multivariate quality control charts.

In this paper, we shall introduce the application study of the most common using technique of multivariate quality control chart; T^2 chart . In addition, this study suggests an improvement of T chart by using PCA are called T^2 -PC chart. we used principal component for plotting (T^2 -PC) chart by depend on MATLAB and MINITAB. The results of the following control charts compared with each other.

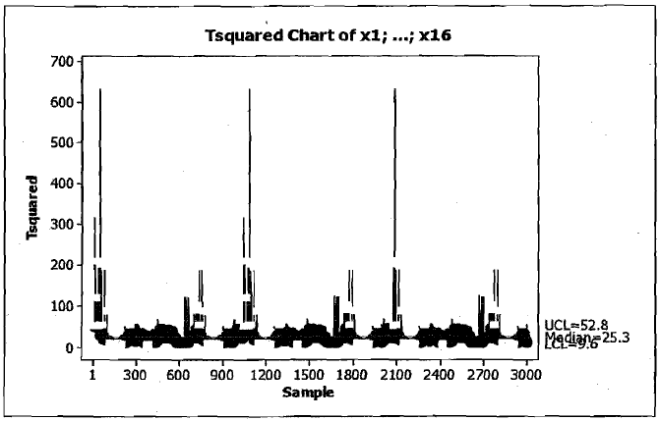

Fig. (2) T^2 chart of x_1, ……., x_{16}

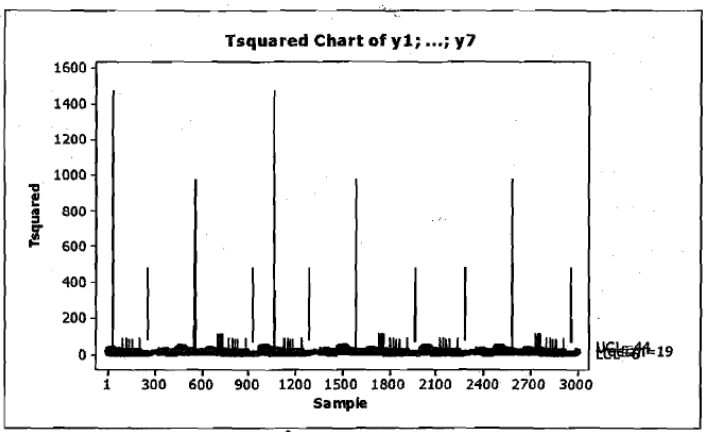

Fig. (3) T^2 chart of y_1, ……., y_7

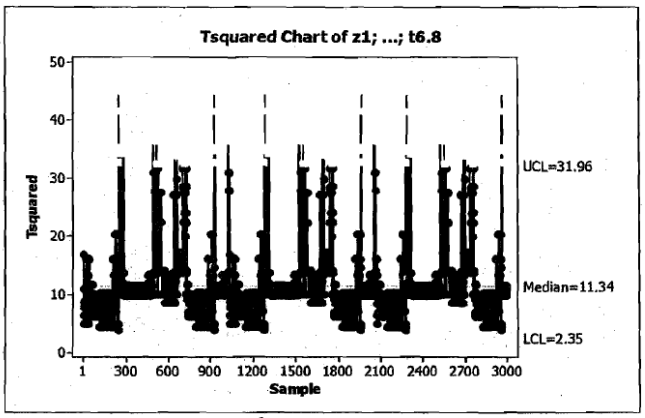

Fig. (4) T^2 chart of z_1, ……., $t_{6.8}$

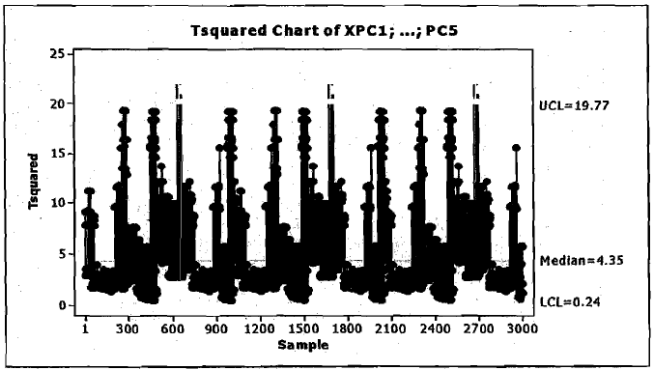

Fig. (5) T^2 chart of X PC1, ……., PC5

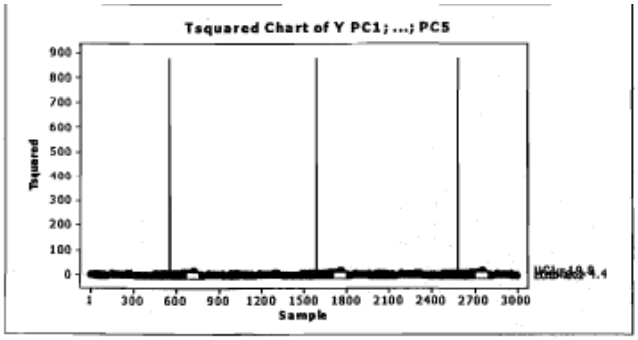

Fig. (6) T^2 chart of y PC1, ……., PC5

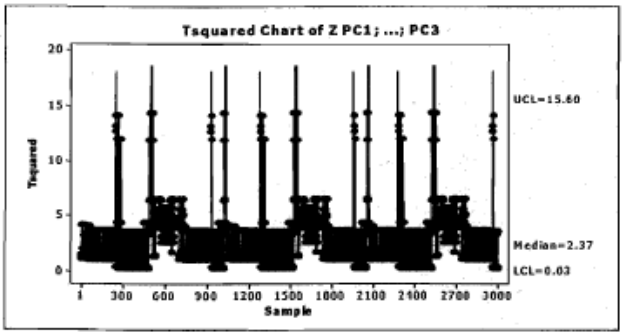

Fig. (7) T^2 chart of z PC1, ……., PC3

Results of T^2 chart and T^2-PC chart

High pressure stage

T^2 chart indicate that the out-of-control rate 15.76 % and the in-control rate 84.24%, while the test result of T^2-PC chart indicate that the out-of-control rate 1.267% and the in-control rate 98.733%.

Low-pressure stage

T^2 chart indicate that the out-of-control rate 2.59% and the in-control rate 97.41%, while test result of T^2-PC chare indicate that the out-of-control rate 2.2% and the in-control rate 97.8%.

Evaporation and Prilling stage

T^2 chart indicates that the out-of-control rate 2.87% and the in-control rate 97.13%, while test result in T^2-PC chare indicates that the out-of-control rate 1.27% and the in-control rate 98.73%.

4. Conclusions:

The Multivariate Statistical Process Control -based on Principal Component Analysis (MSPC-PCA) an efficient statistical tool for process understanding and It used to reduce the dimensionality of the monitoring space by projecting the information in the original variables down onto lowdimensional subspace defined by a few latent variables the process is then

monitored in these latent subspace by using a few multivariate control charts built from multivariate statistics which can be thought of as process performance indices .The new latent variable summarize the most important information of the original K variables .

The concept of principal component is shown graphically in figure (1) .The figure shows a three dimensional data set where the data lie primarily in a plane , thus the data is well described by a two principal component (PC) model .The first eigenvectors of PC aligns with the greatest variation in the data while the second PC aligns with the greatest amount of variation that is orthogonal to the first PC.

References

[1] Box, G .E, Hunter, W.G., and Hunter ,J.S. (1978). ”Statistics for Experimenters. New York: Wiley.

[2] Eriksson ,L., Johansson, E., Wold, N .,and Wold , S. (2001)”Multi-and Multivariate Data Analysis Principals and Applications. ”U metrics AB.

[3] Hanaa, M .S ,Abd ELftah ,M .K and Mohamed, S .E (2007)” Statistical Study of Multivariate Quality Control Procedures and Its Application” Master Dissertation, Banha University, Faculty Commerce, Department of Applied Statistics.

[4] Jackson, J.E. (2003). "A User's Guide to Principal Components ". John Wiley & sons, New York.

[5] Jackson, J.E., and Mudholkar, G.S. (1979). "Control Procedures for Residuals Associated With Principal Component Analysis". Technometrics, 21, PP. 341- 349.

[6] Jesus, M.A and Verde, C.A (2007). ” Fault Detection for large scale systems using Dynamic principal component Analysis with Adaptation International Journal of Computers, Communications and control vol. 11(2) pp. 185-194.

[7] Kourti, T (2005).”Application of latent variable methods to process control and multivariate statistical process control in industry .Int. J. Adapt Control Signal process, Vol. 19 pp. 213-246.

[8] Kresta, J.V, and Macgregor, J.E, and Marlin, T.E. (1991). ”Multivariate Statistical Monitoring of process operation performance” .Computing Chemical Engineering, vol.69 .pp.35-47.

[9] Lowry, C.A.; Woodall, W.H.; champ, C.W.; and Rigdon, S. E. (1992). "A Multivariate Exponentially Weighted Moving Average Control Chart". Technometrics, 34, pp. 46-53.

[10] MacGregor, F.J. and kourti, T. (1995). "Statistical process control of multivariate process". Control Eng. Practice. Vol. 3. No. 3, pp. 403-414.

[11] Martens, H., Naes, T.(1989). ”Multivariate Calibration”. New York: Wiley.

[12] Seshu, K.D., and Madhusree, K.U. (2011). "Multivariate Statistical Process Monitoring and Control “Master Dissertation, National Institute of Technology ,Chemical Engineering Department”.

[13] Stella, B.I. and Aggeliki ,K.I. (2008) ”Application of Principal Component Analysis for Monitoring and Disturbance Detection of process”. Engineering Research Institute, vol. 47, pp, 6972-6987.

[14] Tracy, N.D.; Young, J. C.; and Mason, R.L. (1992). "Multivariate control charts for Individual Observations". Journal of Quality Technology. 24, pp. 88-95.

[15] Woodall, W.H., (2000).”Controversies and Contradictions in statistical process control .” Journal of Quality Technology. Vol. 32(4), pp. 341-350.