Multivariate Statistical Process Control Based on Principal Component Analysis: Implementation of Framework in R

Ana Cristina Braga[1], Cláudia Barros[1], Pedro Delgado[2], Cristina Martins[2], Sandra Sousa[1], J. C. Velosa[2], Isabel Delgado[2], and Paulo Sampaio[1]

[1] ALGORITMI Centre, University of Minho,
4710-057 Braga, Portugal
acb@dps.uminho.pt
[2] Bosch Car Multimedia Portugal SA, apartado 2458,
4705-970 Braga, Portugal
Pedro.Delgado@pt.bosch.com

Abstract The interest in multivariate statistical process control (MSPC) has increased as the industrial processes have become more complex.
This paper presents an industrial process involving a plastic part in which, due to the number of correlated variables, the inversion of the covariance matrix becomes impossible, and the classical MSPC cannot be used to identify physical aspects that explain the causes of variation or to increase the knowledge about the process behaviour.
In order to solve this problem, a Multivariate Statistical Process Control based on Principal Component Analysis (MSPC-PCA) approach was used and an R code was developed to implement it according some commercial software used for this purpose, namely the ProMV (c) 2016 from ProSensus, Inc. (www.prosensus.ca).
Based on used dataset, it was possible to illustrated the principles of MSPC-PCA.
This work intends to illustrated the implementation of MSPC-PCA in R step by step, to help the user community of R to be able to perform it.

Keywords: Multivariate Statistical Process Control (MSPC), Principal Component Analysis (PCA), Control Charts, Contribution plots, R language.

1 Introduction

Modern production processes have become more complex and now require a joint analysis of a large number of variables with considerable correlations between them [13].

With univariate statistical process control (SPC), it is possible to recognize the existence of assignable causes of variation and distinguish unstable processes from stable processes where only common causes of variation are present. The main SPC charts are Shewhart, CUSUM and EWMA charts. They are easy to use and enable to discriminate between unstable and stable processes. This

way, it is possible to detect many types of faults and reduce the production of non-conform products [14].

Although SPC Shewhart charts were designed to control a single characteristic, if more than one characteristic is relevant to the process and these characteristics are independent, the use of those charts is still the right choice. However, a separate analysis of correlated variables may lead to erroneous conclusions.

Figure 1 describes a process with two quality variables (y_1,y_2) that follow a bivariate normal distribution and have a $\rho(y_1, y_2)$ correlation. The ellipse represents a contour for the in-control process; the dots represent observations and are also plotted as individual Shewhart charts on y_1 and y_2 vs. time. The analysis of each individual chart shows that the process appears to be in statistical control. However, the true situation is revealed in the multivariate y_1 vs. y_2, where one observation is spotted outside the joint confidence region given by the ellipse [10].

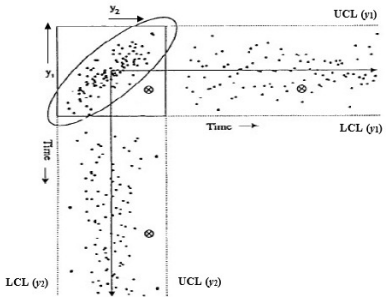


Figure 1: The misleading nature of univariate charts (adapted from [10]).

When applying a multivariate statistical approach for monitoring the status of a process, a set of difficulties can be found. Some of them are listed in [11], as follows:

1. Dimensionality: large amounts of data, including hundreds or even thousands of variables (e.g. chemical industry);
2. Collinearity among the variables;
3. Noise associated with the measurement of process variables;
4. Missing data: the largest data sets contain missing data (sometimes up to 20%).

Thus, it is necessary to find methods to help overcome these difficulties.

In complex processes with a large number of variables (tens, hundreds or even thousands), another problem, associated with collinearity, should be considered: the inversion of the variance/covariance matrix to compute the distance of Hotelling's T^2 becomes difficult or even impossible (singular matrix). In such cases, the traditional multivariate approach must be extended and the principal component analysis (PCA) used in order to obtain new uncorrelated variables. This process is achieved through a spatial rotation, followed by a projection of the original data onto orthonormal subspaces [7].

The R language provides a flexible computational framework for statistical data analysis. R has several packages and functions to perform the PCA, and a recent one to perform the multivariate statistical quality control (MSQC) [18], but the sequence to perform MSPC-PCA is missing and hard to follow.

This study describes an R code that covers all the main steps of the MSPC-PCA in an industrial context. All computation implemented in R follow the procedures used by ProSensus Commercial Software, which deals with multivariate data analysis for a large number of variables.

The main packages used in this study were `prcomp`, `psych`, `FactoMineR` or `pcaMethods`.

2 Multivariate Statistical Process Control Based on PCA

The use of PCA aims to reduce the dimensionality of a dataset with a large number of correlated variables by projecting them onto a subspace with reduced dimensionality [8]. These new variables, the principal components (PCs), are orthogonal and can be obtained through a linear combination of the original variables [3].

Multivariate control charts based on the PCA approach provide powerful tools for detecting out of control situations or diagnosing assignable causes of variation. This function was illustrated by monitoring the properties of a low-density polyethylene produced in a multi-zone tubular reactor, as presented in [10].

2.1 Principal Components, Scores and Loadings

To perform PCA, considering a data set given by a matrix $\boldsymbol{X}$, where n and p are, respectively, the number of observations (rows) and the process variables (columns). As a process can have different variables expressed in different units, before applying PCA, the variables are usually standardized by scaling them to zero mean and unit variance. The packages `prcomp`, `pcaMethods` (available in Bioconductor) and `FactoMineR` performs this kind of analysis.

2.2 Representation of the Observations in the Reduced Dimension PCA Model - Geometric Interpretation

The equation $\boldsymbol{T} = \boldsymbol{P}^{'}\boldsymbol{X}$ is interpreted as a rotation of the axis system composed of the original variables X into a new axis system composed of the PCs.

As mentioned earlier, most of the variability in the original data is captured in the first m PCs. Thus, the previous equation for the full PCA model can be written for a new reduced dimension model [14]:

$$\boldsymbol{T_m} = \boldsymbol{P_m^{'}}\boldsymbol{X} \Rightarrow \boldsymbol{X} = \boldsymbol{T}_m\boldsymbol{P}_m^{'} + \boldsymbol{E} = \sum_{i=1}^{m} \boldsymbol{t}_i\boldsymbol{p}_i^{'} + \boldsymbol{E} \tag{1}$$

where $\boldsymbol{E}$ is the residual matrix given by the difference between the original variables and their reconstruction using the reduced dimension PCA model.

The geometric interpretation of the previous equations is the projection of the original variables onto a subspace of dimension $m < p$ after the previously described rotation.

The concept is illustrated in Fig. 2, where a three-dimensional data set is represented, as are its projection (scores) in a plane with two-dimensions (PC1 and PC2) [11].

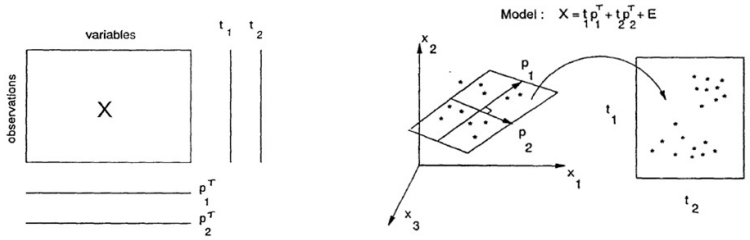


Figure 2: PCA as a data projection method (source: [11]).

Four types of observation can be found with this projection of data:

1. "regular observations": in accordance with the PCA model defined;
2. "good leverage points": close to the PCA subspace but far from the center;
3. "orthogonal outliers": with a long orthogonal distance to the PCA subspace, but close to regular observations, when looking at their projection onto the PCA subspace;
4. "bad leverage points": with a long orthogonal distance and far from the regular observations [4].

2.3 Number of Principal Components

The number m of principal components retained to build the PCA model can be defined by using some of the following methods: the amount of variability explained by the PCA model (R^2), the Kaiser method, the scree plot, the broken stick or the cross-validation (Q^2) [8]. When used individually, none of these methods is definitive. Some commercial software packages specialized in MSPC, such as ProMV from ProSensus, use a joint analysis of R^2 and Q^2.

The percentage of variability (R^2) explained by the model is directly related to the number of principal components considered for the PCA model and can be computed by $100 \times (\sum_i^m \lambda_i / \sum_i^p \lambda_i)\,\%$, where λ_i corresponds to the eigenvalue for PCi [8].

The cross-validation (Q^2) describes the predictive ability of the proposed model and is based on the evaluation of prediction errors of the observations not used to build the model [21]. For the training data, the prediction error decreases as more components are added. However, for the testing data, i.e., observations that were not used to build the model, this error increases when too many components are used. This effect happens because the model is being

over-fitted with noise. The number of components to be considered is the one with the smallest prediction error (Fig. 3).

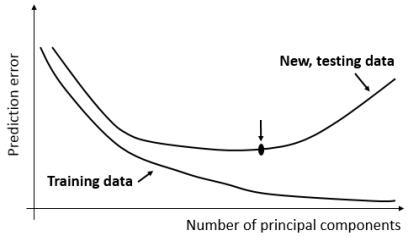


Figure 3: Number of components in the model: joint analysis of R^2 and Q^2 (adapted from [2])

2.4 Multivariate Control Charts Based on PCA for Detecting Assignable Causes

Take in to account that T^2 statistic is the weighted distance of an observation to the center of the PCA subspace, the weighting factor is the variation in the direction of the observation so T_m^2 can be computed as follows [10]:

$$T_m^2 = \sum_{i=1}^{m} \frac{\boldsymbol{t}_i^2}{s_{t_i}^2} = \sum_{i=1}^{m} \frac{\boldsymbol{t}_i^2}{\lambda_i} \tag{2}$$

The upper control limit for T^2, with $100(1-\alpha)\%$ confidence, is given by the F-distribution with m and $n-m$ degrees of freedom [10]:

$$UCL\left(T_m^2\right) = \frac{m\left(n^2-1\right)}{n\left(n-m\right)} F_{\alpha,m,n-m} \tag{3}$$

It can also be approximated by the chi-square distribution [15]:

$$UCL\left(T_m^2\right) = \chi_{m,\alpha}^2 \tag{4}$$

The square prediction error (SPE) or Q statistics is related to the variability in the PCA model and can be defined as the quadratic orthogonal distance [10]:

$$SPE = \sum_{j=1}^{p} \left(\boldsymbol{x}_{new,j} - \hat{\boldsymbol{x}}_{new,j}\right)^2 \tag{5}$$

Assuming that residuals follow a multivariate normal distribution, the upper control limit for the SPE chart can be computed using the following equation [6]:

$$UCL\left(SPE_\alpha\right) = \theta_1 \left[\frac{z_\alpha \sqrt{2\theta_2 h_0^2}}{\theta_1} + \frac{\theta_2 h_0 \left(h_0 - 1\right)}{\theta_1^2} + 1 \right]^{1/h_0} \tag{6}$$

where, $h_0 = 1 - \frac{2\theta_1\theta_3}{3\theta_2^2}$, $\theta_i = \sum_{j=m+1}^{p} \lambda_j^i$ with $i = 1, 2, 3$ and z_α is the value of the standard normal distribution with level of significance α.

According to [17], an approximation of SPE, based on a weighted chi-square distribution, can be used, as follow:

$$UCL\left(SPE_{\alpha}\right) = \frac{\nu}{2b}\chi^{2}_{\frac{2b^{2}}{\nu},\alpha} \tag{7}$$

where b is the sample mean and ν is the variance.

2.5 Diagnosing Assignable Causes

After detecting a faulty observation, the PCA model should be able to identify which variables contribute most to this situation.

Contribution plots were firstly introduced by [12] and decompose the fault detection statistics into a sum of terms associated with each original variable. Consequently, the variables associated with the fault should present larger contributions. This way, using contribution plots, it is possible to focus the attention on a small subset of variables, thus making engineer and operator diagnostic activities easier [9].

As there is no unique way to decompose these statistics, various authors have proposed different formulas to calculate the contributions [9]. Westerhuis et al. [20] discussed the contribution plots for both the T^2 and SPE statistics in the multivariate statistical process control of batch processes. In particular, the contributions of process variables to the T^2 are generalized to any type of latent variable model with or without orthogonality constraints. Alcala and Qin [1] assigned these contributions to three general methods: complete-decomposition, partial-decomposition and reconstruction-based contributions.

The contribution to T^2 of a variable x_j, for m PCs, is given by:

$$cont_{j}^{T^{2}} = \boldsymbol{x_j}\sqrt{\sum_{i=1}^{m}\left(\frac{\boldsymbol{t_i}}{\boldsymbol{s}_{ti}}\right)^{2}\boldsymbol{p}_{i}^{2}} \tag{8}$$

The contribution to SPE of a variable x_j, for m retained PCs, is given by:

$$cont_{j}^{SPE} = e_{j}^{2} \times sign(e_{j}) \tag{9}$$

where $e_j = x_j - \hat{x}_j = x_j - \sum_{i=1}^{m} t_i p_i$

2.6 Steps for Applying MSPC-PCA

To apply the MSPC-PCA it is necessary to follow the following steps:

(1) Collection of a sample representative of the normal operating conditions (NOC);
(2) Application of PCA: use of `prcomp` function in R, the standardization is included;

(3) Definition of the number of principal components to be retained: the `FactoMineR` package could be used to produce the same results of `prcomp` and we can chose directly the number of components as parameter in the function. Another to perform PCA is `pcaMethods` that uses some measures for internal cross validation techniques;
(4) Interpretation of the model obtained: analysis of scores and loadings plots. To draw these graphs we use the package `ellipse` and `plot`;
(5) Identify the physical meaning of each of the principal components, if existing;
(6) Plot control charts for T^2 and SPE defining the limits according the equations;
(7) Interpretation of contributions plot and elimination of strong outliers.

3 Results

This section will present the scripts of R code for the R user community to be able to perform MSPC-PCA by following all the necessary steps described in section 2.6. For each step an example of a dataset of a plastic part will be presented. The goal of this study was to identify which geometrical dimensions of this plastic parts had the highest variability.

All calculation methods used were implemented in R programming language. The most important packages and sections of the R codes were included for reference.

The plastic parts used in this study were selected from the same production batch on three different days (20 parts per day). The mold had two cavities and 86 geometrical dimensions, such as flatness, length, width and thickness, which were measured with a coordinate measuring machine. This dataset will be designated, in the R code, by `dataset`.

3.1 Model Summary

PCA is aimed to produce a small set of independent principal components, from a large set of correlated original variables. Usually, a smaller number of PCs explains the most relevant parts of variability in the data set. The method used to decide the number of PCs to retain was the joint observation of two indicators: R^2, which is a quantification of the explained percentage of variation obtained directly with the eigenvalues; and Q^2, which measures the predictive ability of the model and is obtained through cross-validation.

The R^2 can be computed by using the function `prcomp` included in the `stats` package of R, as follows:

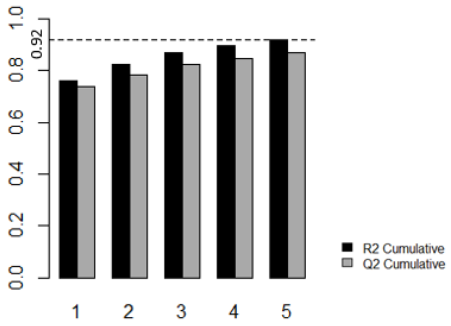
```
acp<-prcomp(dataset,scale=T)
```

The `FactoMineR` package also provides a list of results for multivariate analysis methods, such as PCA, correspondence analysis or clustering [5].

In this work, cross-validation was obtained with the `pcaMethods` package. It provides a set of different PCA implementations, together with tools for cross-validation and visualization of the results [19]. The code used to perform the cross-validation was:

```
pc.Meth.sca.cv<-pca(scale(dataset), nPcs=5, method = 'svd',cv='q2')
plot(pc.Meth.sca.cv)
cv.tab<-as.data.frame(cvstat(pc.Meth.sca.cv))
```

Both indicators, R^2 and Q^2, suggested that five PCs were enough to explain the relevant part of the variability associated to the production process of the plastic part (Fig. 4 and Table 1). The total variation explained by the model with five components was approximately 92%



Comp.	Cumulative R^2	Cumulative Q^2
PC1	0.7605	0.7418
PC2	0.8255	0.7755
PC3	0.8686	0.8043
PC4	0.8977	0.8253
PC5	0.9204	0.8387

Table 1: Tabular result.

Figure 4: Graphical result.

3.2 Score Plots

Score plots are useful graphical analysis tools. Timeline score plots for a single PC are used to analyze time-related variation. Scatter plots of the combination of two PC scores are used to analyze the presence of clusters and how is each observation aligned with each one of the PCs. Observations that lie outside of the control limits may represent outliers. Score plots in this paper showed the control limits for 95% (dashed ellipse) and 99.7% (continuous ellipse).

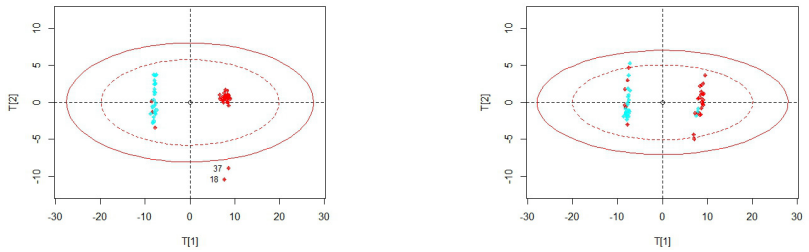
Score plots can be obtained with the function *ellipse*, which creates the outline of a confidence region for two score variables [16]. Part of the R code used to obtain the score plots for PC1 and PC2 is the following:

```
a.acp<-acp$x[,1:2]
centros.acp<-colMeans(a.acp)
lcov.acp=solve(cov(a.acp))
lcov.acp
plot(ellipse(type = "chi",cov(acp$x[,c(1,2)]), level=0.95),
type="l", xlab= "T[1]", ylab="T[2]",col='red', lty=2,xlim=c(-28,28),ylim=c(-12,12)
points(centros.acp [1], centros.acp [2], pch=1)
abline(h=0,lty=2)
abline(v=0,lty=2)
points(ellipse(type = 'chi',cov(acp$x[,c(1,2)]), level=.997), col='red',type="l")
tipo.obs<-substr(abbreviate(amostra[,1]), start = 5,stop = 5)
tipo.obs
cores<- ifelse(tipo.obs=='1' ,"5", "2")
points(-acp$x[,c(1,2)], cex= 0.75, pch=10, col=cores)
text(x = -acp$x[18,1], y = -acp$x[18,2],labels = 18,cex = .8, pos = 2)
text(x = -acp$x[37,1], y = -acp$x[37,2],labels = 37,cex = .8, pos = 2)
```

By analyzing the score plots (Fig. 5), the physical meaning associated with the PCs retained can be identified. PC1 distinguishes two clusters corresponding

to each one of two cavities in the mold; PC2, which explains approximately 6.5% of the variation, is highly influenced by observations 18 and 37, which are outliers. If these two observations are removed and the new model is built, then PC2 becomes influenced by parts warpage, which is present in both cavities and explains approximately 5%.

The percentages of the total variance explained by PC3, PC4 and PC5, 4.1%, 2.5% and 2.4%, reflect different machine adjustments or different machine/raw material conditions in the different production days. However, each one is so small that their effects were not further analyzed.



(a) Score plot with outliers. (b) Score plot without outliers.

Figure 5: Score plot for PC1/PC2.

The code used to perform the time line score (Fig. 6), for example, for PC1 was:

```
n<-nrow(amostra)
n
plot(1:n,-acp$x[,1],ylim = c(-12,12), main = "Time series plots for PC1",xlab = 'OBS',ylab = 'T[1]')
lines(1:nrow(variaveis),-acp$x[,1],type="b",col='black',pch=19)
abline(h=-5.854,lty=1,col='red')
abline(h=-3.781,lty=2,col='red')
abline(h=5.854,lty=1,col='red')
abline(h=3.781,lty=2,col='red')
text(58,5.854, labels='0.997',pos = 3,cex=0.8,col='red')
text(58,3.781, labels='0.95',pos = 3,cex=0.8,col='red')
text(58,-5.854, labels='0.997',pos = 3,cex=0.8,col='red')
text(58,-3.781, labels='0.95',pos = 3,cex=0.8,col='red')
```

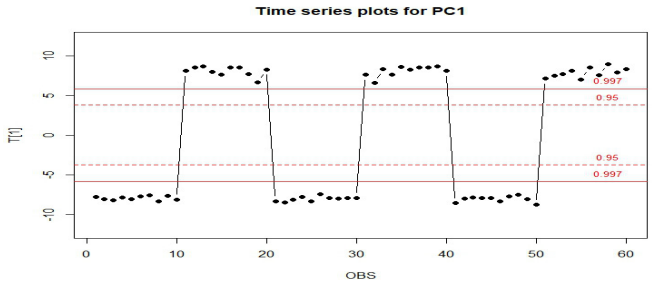


Figure 6: Time series score plot for PC1.

3.3 Loading Plots

Loading plots display the projection of the unit vector with the direction of each original variable in the new PCA axis system. When represented in a scatter plot, the variables that are strongly correlated with a PC create a small angle with this PC direction. The variables that are closer to the center of the plot are not relevant for explaining the variation associated with this PCs pair.

Part of the R code used to compute the loadings plot for PC1 and PC2 is:

```
load<-sweep(pca3$var$coord,2,sqrt(pca3$eig[1:ncol(pca3$var$coord),1]))[,1:ncol(pca3$var$coord)]
plot(load.stand[,c(1,2)], xlim=c(-.2,.2),ylim=c(-.40,.40),xlab='PC1',ylab = 'PC2')
abline(h=0,lty=2)
abline(v=0,lty=2)
text(load.stand[,1],load.stand[,2],labels =colnames(dataset),cex=0.8, lwd=2,col="blue")
```

The loadings plot $PC1 - PC2$, in Fig. 7, show the variables that are positively or negatively correlated with $PC1$, which is already known to represent the different cavities. The variables with high loadings in the $PC2$ are describing warpage, as already mentioned.

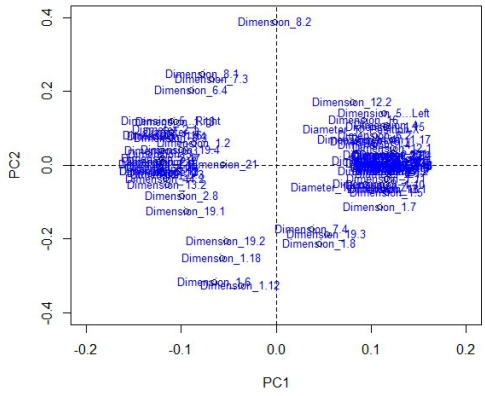


Figure 7: Loadings plot $PC1 - PC2$.

3.4 Hotelling's T^2 control charts and contributions plot

T^2 indicates the distance from an observation to the center of the PCA subspace; it is a summary statistics calculated as the sum of squares of the scores of each observation in each one of the retained principal components. In the case of plotting only two-dimensions, all points on in the ellipse have the same T^2 value and correspond to an upper limit (95% or 99.7%) estimated from the model.

The R code used for T^2 is the shown below and follows the eq.2.

```
num.com <- 5
a.acp <- acp$x[,1:num.com]
centros.acp <- colMeans(a.acp)
```

```
lcov.acp = solve(cov(a.acp))
dm.acp <- rep(0,length(a.acp[,1]))
for(i in 1:length(a.acp[,1])){
dm.acp[i]=round(t(a.acp[i,]-centros.acp)%*%lcov.acp%*%(a.acp[i,]-centros.acp),3)
}
```

The upper limits (95% and 99.7%), according to eq. 3, are computed by using the following code:

```
k<-num.com
n<-nrow(dataset)
cc.sw.UCL.997<-(k*(n+1)*(n-1))/(n*(n-k)) * qf(.997,k, n-k)
cc.sw.UCL.95<-(k*(n+1)*(n-1))/(n*(n-k)) * qf(0.95,k, n-k)
```

Using the dataset of a plastic part, the control charts for T^2 shown in Fig. 8 suggest that assignable causes of variation were associated with observations 18 and 37. Thus, a method that allows identifying the original variables associated with these assignable causes of variation is required. This method is the calculation of contributions.

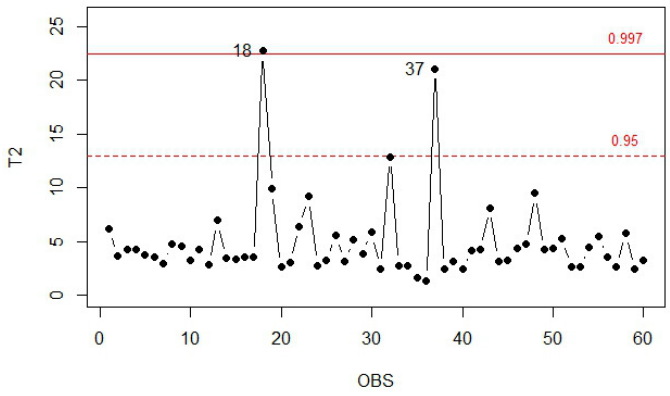


Figure 8: Hotelling's T^2 control chart.

Contribution plots are used to identify which variables contribute more to T^2 values. The observations 18 is further analyzed concerning their contributions to T^2. Since the effect observed in the observation 18 is the same as in the observation 37, the contribution plots for the observation 37 were not presented in this study. Part of the code used to perform the contribution plot to T^2 is (based on eq. 8):

```
data <- matrix(NA, nrow=num.com, ncol=ncol(dataset))
for (i in 1:num.com){
num=round(t(a.acp[1,i]-centros.acp[i])%*%lcov.acp[i,i]%*%(a.acp[1,i]-centros.acp[i]),3)
data[i,]<-num%*%load[,i]^2
}
contr<-sqrt(colSums(data))*scale(dataset)[1,]
barplot(contrP,axes = T, ylim=c(-8,7),cex.axis = .9)
```

Considering the previous analyses related to loadings and scores, and the high contributions of the variables Dimension 1.12, Dimension 1.6, Dimension 8.2 and Dimension 7.3 and Dimension 8.1, as shown in Fig. 9, it can be concluded that this part has a problem of planeness (Dimensions 1.6 and 1.12) associated with a reduced thickness (Dimensions 8.1 and 8.2).

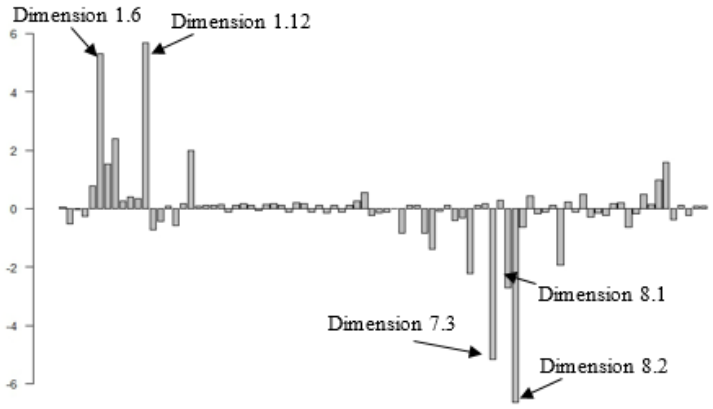


Figure 9: Contribution Plot to Hotelling's T^2 for the observation 18.

3.5 *SPE* control charts and contributions plot

Observations with high SPE show that some of the variables varied in a different direction from what was expected, considering the correlation structure of the original variables. In other words, T^2 measures the distance to the center of the model (many variables are far from their average values without breaking the correlation structure) and the SPE measures the distance to the model (correlation structure strongly broken).

Part of the R code used to compute SPE is shown below and follows the eq. 5:

```
num.com.spe<-as.numeric(5)
a<-pca3$ind$coord[,1:num.com.spe]
load<-sweep(pca3$var$coord,2,sqrt(pca3$eig[1:ncol(pca3$var$coord),1]),FUN="/")[,1:num.com.spe]
Ye<-a %*% t(load)
Qt<-rowSums((scale(dataset)-Ye)^2)
```

According to **ref13**, the R code that should be used to perform the upper limits (95% and 99.7%) for SPE is the following (based on eq. 6):

```
QCL99.7<-(var(Qt)/(2*mean(Qt)))*qchisq(p = .997,df = (2*mean(Qt)^2)/var(Qt))
QCL95<-(var(Qt)/(2*mean(Qt)))*qchisq(p = 0.95,df = (2*mean(Qt)^2)/var(Qt))
```

The control charts for SPE is illustrated in Fig. and suggest that assignable causes of variation were associated with observation 28.

To compute the contributions to SPE and the respective chart, the following code can be used (based on eq. 9):

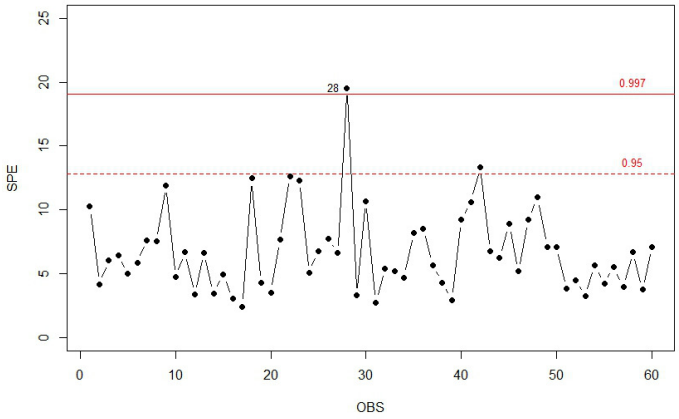


Figure 10: SPE control chart.

```
a<-pca3$ind$coord
load<-sweep(pca3$var$coord,2,sqrt(pca3$eig[1:ncol(pca3$var$coord),1]),FUN="/")
Ye<-a %*% t(load)
erros<-scale(variaveis)-Ye
CONT<- matrix(NA, nrow= nrow(dataset), ncol=ncol(dataset))
for(i in 1:nrow(erros)){
CONT[i,]<-sign(erros[i,])*erros[i,]*t(erros[i,])}
barplot(CONT[28,],axes = T,names.arg=colnames(dataset),ylim=c(-4,6),cex.axis = .9)
```

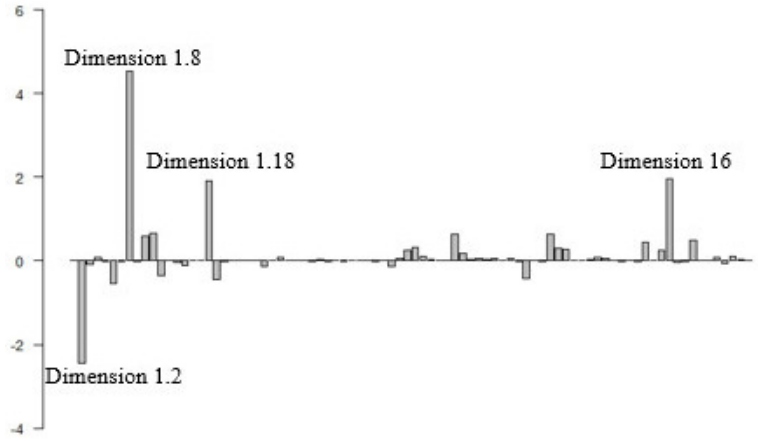


Figure 11: Contributions Plot to SPE for the observation 28.

In Fig. 11, the variable Dimension 1.8, with a high contribution to SPE, does not show the usual variability associated with different cavities. In this

observation, Dimension 1.8 has an increased value without being followed by other variables that should have a high correlation with it.

3.6 Raw Data: The original values of data set

Once the variables that contribute most to T^2 and SPE are detected, the original variables are analyzed to check which could contribute to a product malfunction.

Regarding the variables with high contributions in the observation 18, raw data of the original variables show that they have values lower than what would be expected, as presented in Fig. 11 for the variable Dimension 8.1. It was previously referred that the observation 37 had the same effect as the observation 18; this conclusion is confirmed by the raw data of the variable Dimension 8.1, presented in Fig.12.

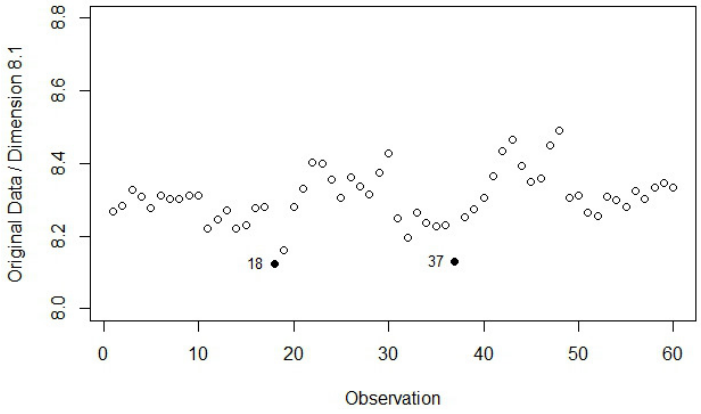


Figure 12: Original values for variable Dimension 8.1.

4 Conclusions

In this paper, an overview of the main MSPC-PCA concepts is presented. The application of these methods allows finding a reduced number of new variables that are linear combinations of the original variables. With this reduced number of PCs, a model that explains the most relevant part of the variability can be created and used to control the process.

The main procedures used to implement this MSPC-PCA analysis are the following: computing eigenvectors and eigenvalues; choosing the number of components of the model; calculating score and loading plots; calculating T^2 and SPE control charts. Thus, assignable causes of variation can be detected and the original variables involved can be identified through the calculation of contributions.

These results show that the MSPC-PCA can detect outliers, identify physical aspects that explain causes of variability and analyze the stability of the production process (injection molding). PCA also enables operators and process engineers to increase their knowledge about the way the process behaves and to identify the underlying factors which govern the variability of the process.

The use of the R programming language in an industrial example demonstrates the great potential of MSPC-PCA techniques in multivariate data analysis and multivariate statistical control of processes.

References

1. Alcala, C. F. and Qin,S. Joe: Analysis and generalization of fault diagnosis methods for process monitoring. Journal of Process Control, vol. 21, 322-330 (2011).
2. Aptula, A. O., Jeliazkovab, N. G., Schultzc, T. W. and Cronina, M.T.D.: The Better Predictive Model: High q2 for the Training Set or Low Root Mean Square Error of Prediction for the Test Set?. QSAR and Combinatorial Science, vol. 24, 3, 385-396 (2005).
3. Bharati, M. H. and MacGregor, John F.: Multivariate Image Analysis for Real-Time Process Monitoring and Control. Industrial & Engineering Chemistry Research, vol. 37, 4715-4724 (1998).
4. Hubert, M., Rousseeuw, P. J. and Branden, K. V. ROBPCA: A New Approach to Robust. American Statistical Association and the American Society for Quality, vol. 47, 1,(2005).
5. Husson, F., Josse, J., Le, S. and Mazet, J.: Multivariate Exploratory Data Analysis and Data Mining. CRAN, nov. 2016.
6. Jackson, J. E.: Principal Components and Factor Analysis: Part I - Principal Components. Journal of Quality Technology, vol.14, 11, 201-213 (1980).
7. Jackson, J. E.(1991): A users guide to Principal Components. John Wiley & Sons, New York.
8. Jolliffe, I. T. (1986): Principal Component Analysis. Springer, New York.
9. Kerkhof, P. Van den, Vanlaer, J. , Gins, G and Impe, J. F. M. Van: Analysis of smearing-out in contribution plot based fault isolation for Statistical Process Control. Chemical Engineering Science, vol. 104, 285-293 (2013).
10. Kourti, T. and MacGregor, John F.: Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemometrics and Intelligent Laboratory Systems, vol. 28, 1, 3-21 (1995).
11. MacGregor, John F.: Using On-Line Process Data to Improve Quality: Challenges for Statisticians. International Statistical Review (1997).
12. MacGregor, J., Jaeckle, C., Kiparissides, C.and Koutoudi, M.: Process monitoring and and diagnosis by multi block pls methods. AIChE Journal, vol. 40, 5, 826838 (1994).
13. MacGregor, John F., Yu, Honglu and Salvador Garca Muoz and Jesus Flores-Cerrillo: Data-based Latent Variable Methods for Process Analysis, Monitoring and Control. Computer & Chemical Engineering (2005).
14. Martin, E. B., Morris, A. J. and Zhang, J. Z.: Multivariate statistical process control charts and the problem of interpretation: A short overview and some applications in industry. System Engineering for Automation (1996).
15. Montgomery, D. C. (2009): Introduction to Statistical Quality Control. John Wiley & Sons, New York.

16. Murdoch, D., Chow, E. D., Celayeta, J. M. F.: Functions for drawing ellipses and ellipse-like confidence regions. CRAN, april 2013.
17. Nomikos, P. and MacGregor, John F.: Multivariate SPC Charts for Monitoring Batch Processes. Technometrics, vol.31, 1, 41-59 (1995).
18. Santos-Fernandez, E.: Multivariate Statistical Quality Control Using R. Springer, 14, (2013).
19. Stacklies, W., Redestig, H. and Wright, K: A collection of PCA methods. CRAN, February 2017.
20. Westerhuis, J. A., Gurden, S. P. and Smilde, A. K.: Generalized contribution plots in multivariate statistical process monitoring. Chemometrics and Intelligent Laboratory Systems, vol. 51, 95-114 (2000).
21. Wold, S.: Cross-validatory estimation of the number of components in factor and principal components models. Technometrics, vol. 20, 4, 397-405, (1978).