Robust PCA and Normal Region in Multivariate Statistical Process Monitoring

Jiageng Chen, Jose A. Bandoni, Jose A. Romagnoli
ICI Laboratory of Process Systems Engineering, Dept. of Chemical Engineering,
University of Sydney, NSW, 2006, Australia

The on-line monitoring and evaluation of process operating performance are essential to safe operation and consistent high quality production. This can be achieved using multivariate statistical approaches (MSA). The basic philosophy of MSA is that the behavior of the process is characterized by data obtained when the process is operating well or in *normal region*. Subsequently, future unusual events can be detected by referring the measured process behavior against this *normal region* model, i.e., *normal region* is used here as a calibration criteria of the process.

The successful implementations of MSA have been extensively reported in the literature. A recent review can be found in MacGregor and Kourti (1995). The procedure of MSA is as follows: (1) data representing normal process behavior are collected; (2) multivariate statistical methods, such as principal component analysis (PCA) and partial least squares (PLS), are utilized to compress the data and to extract the information projecting the data into a low dimension space that summarizes all the important information; (3) *normal region* or control chart is configured to monitor the process; (4) diagnosis and identification of the fault sources, if anyone causes the process out of the *normal region*.

In this work, we focus on steps 2 and 3. The first motivation of the current work is to reliably absorb information despite the existence of outliers. The second is to define *normal region* via the data. This has the advantage that no *a priori* assumption has to be made. The use of a robust PCA via project pursuit in place of the classical PCA for modeling normal process behavior, is proposed, and a kernel approach is suggested as an alternative method to define *normal region*.

Robust PCA via Projection Pursuit (PP)

In MSA, PCA is utilized to build up an empirical model to characterize the normal operation of the process. The process measurements are always subject to random and outliers/gross errors. Outliers are usually influential observations, that is, their deletion often causes major changes in estimates, confidence regions, tests, and so on. Classical PCA, like most commonly used statistical techniques, is excessively sensitive to outliers, that is, sometimes a principal component might be created by just the presence of one or two outliers (Xie et al., 1993). Thus, if outliers exist, the coordinate axes of the principal component space might be misdetermined by classical PCA, and reliable statistical process monitoring would not be obtainable based on the nonrepresentative subspace of the original variable space. There are two possible ways to deal with this problem: (1) detect and eliminate the outliers before using classical PCA, or (2) eliminate the influence of the outliers (property of robustness). In this note we take the second way and use the projection pursuit (Huber, 1985) approach. Projection pursuit (PP) techniques search for a lower-dimension subspace such that the configuration of the data obtained in this subspace can reflect the structure and features of the original high-dimension data in an optimal way. One of the valuable characteristics of PP is that it is easy to robustify.

Assuming that the original dimension is d, and defining the projection index (function) as $J(Z)$, which describes the dispersion of a one-dimension projection $\boldsymbol{Z}$, then a set of selected eigenvectors $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_q$, will satisfy

$$\lambda_1 = \max J(\boldsymbol{Z}) = \max J(\boldsymbol{p}^T\boldsymbol{X}) = J(\boldsymbol{p}_1^T\boldsymbol{X}), \quad \|\boldsymbol{p}_1\| = 1$$

$$\lambda_2 = \max J(\boldsymbol{Z}) = \max J(\boldsymbol{p}^T\boldsymbol{X}) = J(\boldsymbol{p}_2^T\boldsymbol{X}), \quad \|\boldsymbol{p}_2\| = 1 \quad \text{and} \quad \boldsymbol{p}_2 \perp \boldsymbol{p}_1$$

$$\lambda_q = \max J(\boldsymbol{Z}) = \max J(\boldsymbol{p}^T\boldsymbol{X}) = J\left(\boldsymbol{p}_q^T\boldsymbol{X}\right), \quad \|\boldsymbol{p}_q\| = 1 \quad \text{and} \quad \boldsymbol{p}_q \perp \boldsymbol{p}_{q-1} \perp \ldots \boldsymbol{p}_1$$

If λ_{q+1} $(q+1 \le d)$ is small enough to be ignored, this suggests that the data spreads mainly on the q-dimension subspace by $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_q$ and the dimension can thus be reduced.

A well-known robust estimator of scale, the median of absolute distance from the sample median, is taken as the one-dimensional projective index

Correspondence concerning this article should be addressed to J. A. Romagnoli.
Current address of J. Chen and J. A. Bandoni: PLAPIQUI (UNS-CONICET), 12 de Octubre 1842, 8000 Bahia Blanca, Argentina.

$$J = \left(\frac{\text{median}(|z_i - z^*|)}{0.6745} \right)^2 \tag{1}$$

where, $z_i = \boldsymbol{p}_i^T \boldsymbol{X}$, z^* is the median of the one-dimension projection, and 0.6745 is a correction factor which makes the estimate of scale asymptotically unbiased when the underlying distribution is normal. Accordingly, Huber-type weights (Li and Chen, 1985) are used for a refining cycle for z^* and J. The basic structure of the robust principal component analysis algorithm is given in the Appendix A.

Kernel Approach to Define Normal Region

Conventional methods of defining *normal region* or control limit is based on the independent, identical distributed (iid) normal deviates assumption on measurement variables or the latent variables. In some industrial implementation, as shown by Nomikos and MacGregor (1994), the central limit theorem often makes the multivariate assumption on the scores reasonable even if they are not iid. However, the iid assumption has often been questioned since successive sets of measurements are not necessarily independent samples of the process variables and thus the central limit theorem cannot strictly be applied (Crowe, 1996).

While no assumption about the underlying probability distribution of the measurements is possible in order to define *normal region*, an assumption is needed in order to apply statistical tests to the results. We propose the use of a density estimation, a kernel approach, as a method for developing *normal region* directly from historical data of the process.

The probability density gives a natural description of the distribution of x. For example, normal distribution $x \sim (\mu, \sigma^2)$ can be described by the density function $f(x)$ as

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu)^2/2\sigma^2} \tag{2}$$

Density estimation is the construction of an estimate of the density function from the observed data. The kernel method, a popular nonparameter approach, is used in this work. The multivariate product kernel estimator could be constructed, based on random sample $X_1, X_2, \ldots, X_n$ from a density f (Scott, 1992)

$$\hat{f}(X) = \frac{1}{nh_1h_2\cdots h_d} \sum_{i=1}^{n} \prod_{j=1}^{d} K\left(\frac{X_i - X_{ij}}{hj} \right) \tag{3}$$

where h is the window width, also called the smoothing parameter or bandwidth; K is known as a kernel function which satisfies the condition

$$\int_{-\infty}^{\infty} K(x)dx = 1 \tag{4}$$

The quality of a density estimate is now widely recognized to be primarily determined by the choice of smoothing parameter h, and only in a minor way by choice of kernel (Scott, 1992; Bowman, 1984). Usually, although not always, K will be a symmetric probability density function, such as the normal density. In our approach, least-squares cross-validation (LSCV) (Bowman, 1984; Rudemo, 1982) is used to select h. The basic idea of LSCV is to construct an estimate from the data and then to minimize the integrated squared error of this estimate over h to give the optimal choice of window width h, that is

$$\text{find } \hat{h} = \underset{h}{\text{Minimize}}\ CV(h) \tag{5}$$

$$CV(h) = \frac{R(K)}{nh} + \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{i \neq j} \{K_h * K_h(X_i - X_j) - 2K_h(X_i - X_j)\}$$

where K is a kernel function; $K_h(x) = K(x/h)/h$; $R(g) = \int g^2(x)dx$; $f * g$ is used to denote the convolution of two functions f and g; $f * g = \int f(u)g(x-u)du$.

Depending on the requirement of confident level (user specified), a corresponding contour can be found and this contour is *normal region* (assuming the value of the contour is f_{constant}). Aiming at process monitoring, one just needs to calculate $\hat{f}(z^*)$ when new process data z^* is coming, and then compare it with f_{constant}. If $\hat{f}(z^*) \geq f_{\text{constant}}$, that indicates the new projected data point (latent variables) is within the *normal region*; otherwise the process is under fault conditions, and further actions are needed, such as fault isolation and detection. Due to the space limitation, this part is referred to Nomikos and MacGregor (1994) and Raich and Çinar (1995). If the number of latent variables is greater than 2, one can still define *normal region* in the same way. However, one may not be able to visualize it. It is worthwhile mentioning that all density estimators are generalized kernel estimates (Terrell and Scott, 1992). So, the kernel approach can be viewed generally as a strategy using nonparametric density estimation to define *normal region*. It is a data driven strategy, since the proposed method directly extracts information from data themselves. When the underlying distribution is iid normal, similar results such as ones from the conventional approaches can be expected, because basically the density estimator is approximating to the true distribution.

Example

A two CSTR system from Bahri et al. (1995) is used as an example to show the performance of the proposed methods. The simulated steady-state data were generated from a normal disturbance (5%) about a nominal point (which is determined for optimal operation, Bahri et al., 1995). For more reality, a 1% Gaussian noise was added to each measurement. The data would reasonably represent measurements collected from an industrial process at time intervals larger than the process time constant or averages of measurement taken over some time period (MacGregor and Jaeckle, 1994).

Figures 1a and 1b show an example when an outlier exists in the whole data set, the results of performing classical PCA on the CSTR system. The first principal component coordinate axes of the two CSTR system are totally different because of one outlier. Recalling the well-known principal component transforming formula $\boldsymbol{Z} = \boldsymbol{P}^T\boldsymbol{X}$, it is clear that the principal component coordinate axes $\boldsymbol{P}$ is the basis of PCA; any unreliable result of $\boldsymbol{P}$ will affect the result of the PCA as

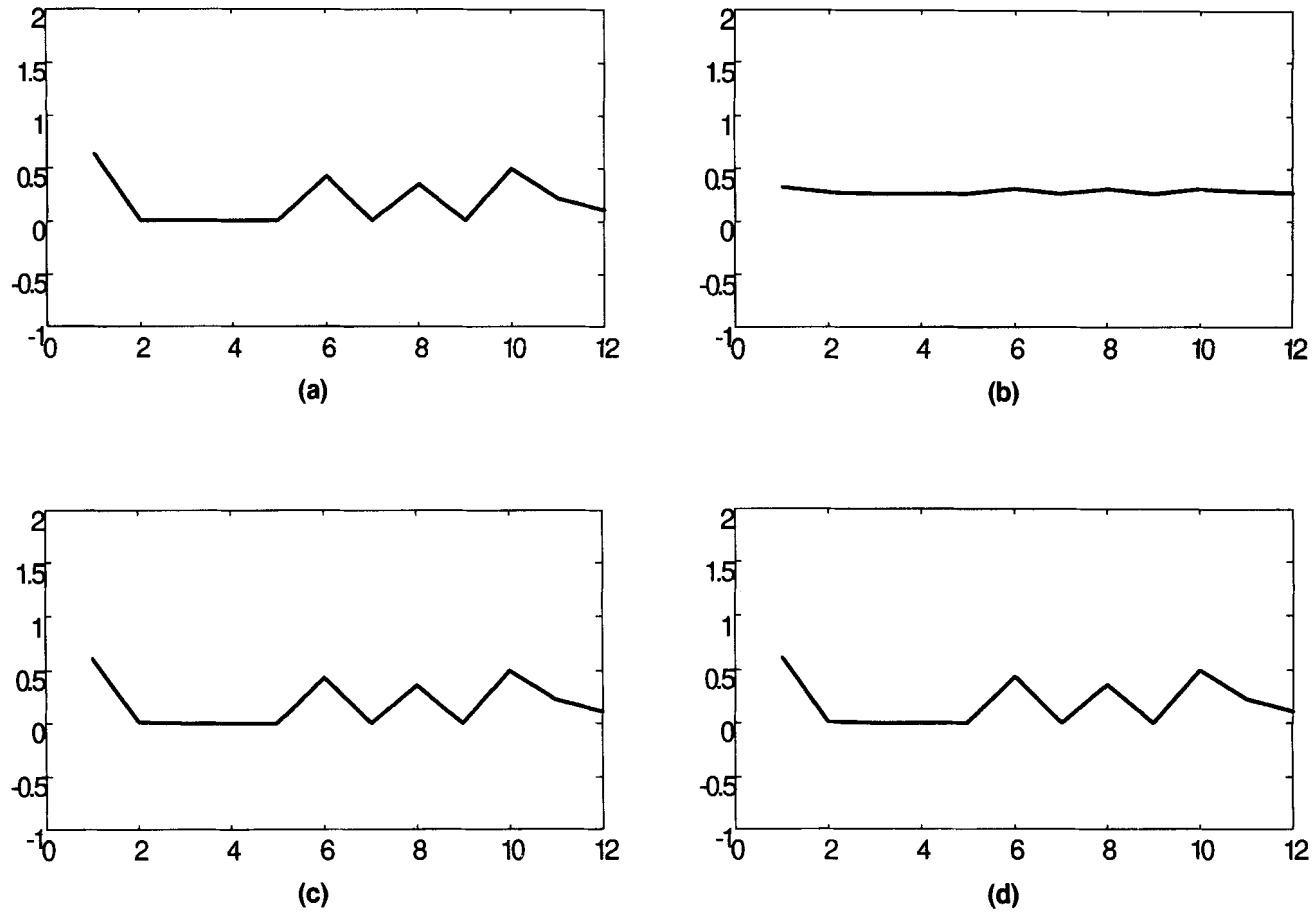

Figure 1. First principal component coordinate axes of the two CSTR system problem.

(a) Classical PCA without outliers; (b) classical PCA with one outlier; (c) robust PCA without outliers; (d) robust PCA with one outlier.

well as the statistical process monitoring. This detriment is eliminated by using robust PCA. As expected, the results of robust PCA, shown in Figures 1c and 1d, are resistant to outliers and yield the same result as classical PCA (without outlier). Conclusively, a reliable result can be expected through using robust PCA, when outliers are present in the reference data. One drawback is that robust PCA requires more computer time.

In the following, we will show the results using the kernel approach for the definition *normal region* for the present example. The number of latent variables was taken equal to two. Figure 2 shows the distribution of latent variables for the two CSTR systems around the operation point. Figure 3 displays the contours of the distribution function. One of the contours could be selected as *normal region* according to the requirement of the confident level. When projected newly

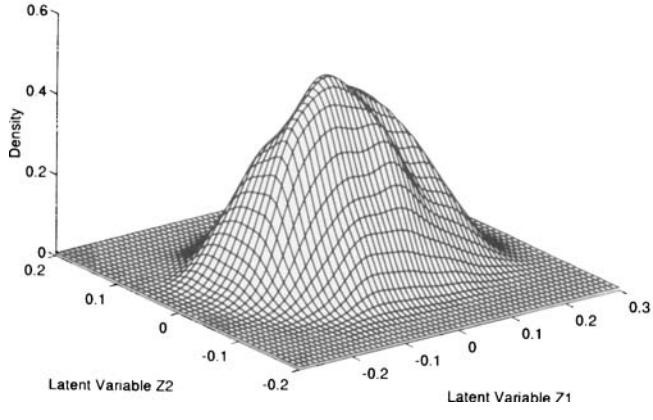

Figure 2. Estimated density function.

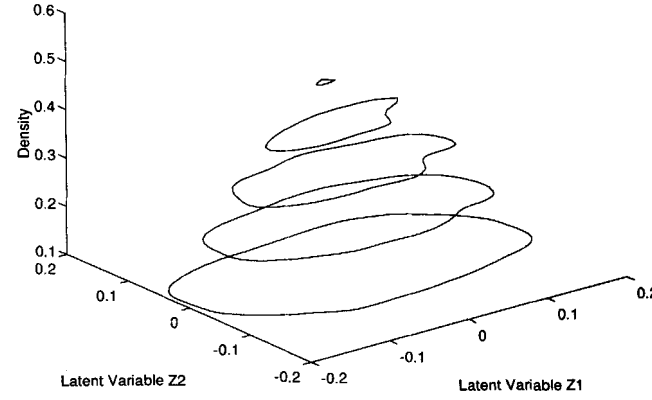

Figure 3. Contours of the density function.

available process data fall into the *normal region*, it indicates the CSTR system is in its normal operating conditions. Otherwise the process conditions have been changed either as results of faults or disturbances, and thus further action should be taken.

Literature Cited

Bahri, P. A., J. A. Bandoni, G. W. Barton, and J. A. Romagnoli, "Back-Off Calculations in Optimising Control: A Dynamic Approach," *Comput. Chem. Eng.*, **19**, S699 (1995).

Bowman, A. W., "An Alternative Method of Cross-Validation for the Smoothing of Density Estimate," *Biometrika*, **76**, 353 (1984).

Crowe, C., "Formulation of Linear Data Reconciliation using Information Theory," *Chem. Eng. Sci.*, **51**(12), 3359 (1996).

Huber, P. J., "Projection Pursuit," *Ann. Stat.*, **13**(2), 435 (1985).

Li, G., and Z. Chen, "Projection Pursuit Approach to Robust Dispersion Matrices and Principal Components: Primary Theory and Monte Carlo," *J. Amer. Stat. Assn.*, **80**(391), 759 (1985).

MacGregor, J., and C. Jaeckle, "Process Monitoring and Diagnosis by Multiblock PLS Methods," *AIChE J.*, **40**(5), 826 (1994).

MacGregor, J. F., and T. Kourti, "Statistical Process Control of Multivariate Processes," *Control Eng. Practice*, **3**(3), 403 (1995).

Nomikos, P., and J. MacGregor, "Monitoring Batch Processes Using Multiway Principal Component Analysis," *AIChE J.*, **40**(8), 1361 (1994).

Raich, A. C., and A. Çinar, "Multivariate Statistical Methods for Monitoring Continuous Processes: Assessment of Discrimination Power of Disturbance Models and Diagnosis of Multiple Disturbances," *Chem. Inte. Lab. Sys.*, **30**, 37 (1995).

Rudemo, M., "Empirical Choice of Histograms and Kernel Density Estimators," *Scand. J. Statist.*, **9**, 65 (1982).

Scott, D. W., *Multivariate Density Estimation: Theory Practice and Visualisation*, Wiley, New York (1992).

Terrell, G. R., and D. W. Scott, "Variable Kernel Density Estimation," *Ann. Statist.*, **20**(3), 1236 (1992).

Xie, Y., J. Wang, Y. Liang, L. Sun, X. Song, and Y. Yu, "Robust Principal Component Analysis by Projection Pursuit," *J. Chem.*, **7**, 527 (1993).

Appendix A: Robust PCA Algorithm

Step 1. Initiation

Assume that q ($q \le d$) principal components need to be found. Initialize the first q principal component directions $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_q$ with the first q columns of a $d \times d$ identity matrix $\boldsymbol{I}$ and let $\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_q)$. Pre-set the projection indices J_i with the median of the original data

$$J_i = \{\text{median}(|x_{ij} - m_i|)\}^2, \quad i = 1, \ldots, d \quad j = 1, \ldots, n$$

$m_i = \text{median}(x_{ij})$; $i = 1, 2, \ldots, d$, $j = 1, 2, \ldots, n$. The initial robust covariance matrix is denoted by

$$\boldsymbol{C} = \boldsymbol{PDP}^T$$

where $\boldsymbol{D} = \text{diag}(J_1, J_2, \ldots, J_q)$.

Step 2. Optimization

Suppose the first $i-1$ principal components $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{i-1}$ and $J_1, J_2, \ldots, J_{i-1}$ are available. The direction with the largest projection index J_i in the orthogonal complement space of the subspace spanned by the $i-1$ selected principal component directions $\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots, \boldsymbol{p}_{i-1}$ should be selected as the ith principal component direction $\boldsymbol{p}_i$. In other words, $\boldsymbol{p}_i$ will be the solution of the following optimal problem

$$\text{Maximize } J = \{\text{median}(|z - z^*|)/0.6745\}^2$$
$$\text{St.:} \quad \|\boldsymbol{p}\| = 1$$

where $z = \boldsymbol{p}^T\boldsymbol{X}$; $\boldsymbol{X}$ is the original data; $\boldsymbol{p}$ is the ith principal component coordinate axes. Then, Huber type iteration (Li and Chen, 1985) is used for a refining cycle for z^* and J

$$z^* = z^* + \sum w_1(t_i)(z_i - z^*) / \sum w_i(t_i)$$
$$J = \sum w_2(t_i)(z_i - z^*)/n$$

where $t_i = (z_i - z^*)/J$, stop criterion is chosen as $|J(\text{new}) - J(\text{old})|/J(\text{old}) \le 0.0001$. After $\boldsymbol{p}_i$ has been determined, one calculates the residual matrix

$$\boldsymbol{X} = (\boldsymbol{I} - \boldsymbol{pp}^T)\boldsymbol{X}$$

In this way the computation is always carried out in the orthogonal complement of the subspace spanned by the already calculated principal component directions, and the resulting principal component direction will be automatically orthogonal to the former directions.

Step 3

After q significant directions have been calculated, the robust covariance matrix is then constructed from them as

$$\boldsymbol{C}^{\text{new}} = \boldsymbol{PDP}^T$$

where $\boldsymbol{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \ldots \boldsymbol{p}_q)$, $\boldsymbol{D} = \text{diag}(J_1, J_2, \ldots, J_q)$.

This modified covariance matrix is compared with the old one. If the difference is smaller than a preset threshold, the computation is terminated (in this research $\|\boldsymbol{C}^{\text{new}} - \boldsymbol{C}^{\text{old}}\| \le 1e-3$); otherwise one returns to Step 2 and a new cycling is invoked by using the estimated principal component directions as initial guesses together with the original data matrix $\boldsymbol{X}$.

Manuscript received Aug. 22, 1995 and revision received May 14, 1996.