Process Monitoring Using Principal Components in Parallel Coordinates

Ricardo Dunia and Thomas F. Edgar
Dept. of Chemical Engineering, The University of Texas at Austin, Austin, TX 78712

Mark Nixon
Emerson Process Management, 12301 Research Blvd., Austin, TX 78759

DOI 10.1002/aic.13846
Published online June 5, 2012 in Wiley Online Library (wileyonlinelibrary.com).

Parallel coordinates is a recognized visualization technique in which data points, each defined by multiple coordinates, are represented by an unlimited number of adjoining parallel axes. This type of visualization technique is suitable for process monitoring applications in industrial facilities where a significant number of sensors are used to detect and identify abnormal operating conditions. This work makes use of principal component monitoring methods implemented in parallel coordinates plots, named PC^2. The PC^2 capabilities to visualize confidence regions of operations, evaluate models with different number of principal components, compare faulty events and determine the frequency of false alarms are here demonstrated. The monitoring visualization technology presented by PC^2 was successfully used for early detection of compressor surge and column flooding using actual process data. © 2012 American Institute of Chemical Engineers *AIChE J,* 59: 445–456, 2013

Keywords: process monitoring, parallel coordinates, principal component analysis, fault detection, multivariable statistical process control, data visualization, PCA

Introduction

Parallel coordinates is a well recognized visualization technique introduced by Inselberg[1,2] in which data points are represented with unlimited number of coordinates by overlaying parallel axes. It is a scalable way to represent data as any additional dimension can be visualized by adding an extra axis without affecting the rest of the graph. Sometimes, for comparison purposes, it is convenient to reorder the axis to group common data features together. Axes can be independently scaled and shifted along for unit conversion or visualization purposes.

Practical applications of parallel coordinates includes pattern recognition,[3] measurement processing for intelligent sensor units,[4] analysis of financial data,[5] enhanced visual analysis of hurricane climate data sets,[6,7] detection of unknown large-scale Internet attacks[8,9] and gene expression data analysis.[10] In the area of monitoring, parallel coordinates have been successfully implemented to monitor temperature variance of a data center[11] and for a wastewater treatment plant.[12]

The use of parallel coordinates may produce visual line clutter due to excessive superposition of lines. Furthermore, the plotting order may provide a different perspective of the data when different colors are used to distinguish data sets. Lun and Zhuo[13] have suggested the use of visual clustering and filtering to diminish the visual cluttering of data lines in parallel coordinate plots. Zhou et al.[14] exploited the use of curve edges to cluster data for more effective data visualization. In this manner, curved lines are used to form visual bundles enhanced by an optimization framework to form clusters. Artero et al.[15] reduced line cluttering by filtering data based on their frequency, and Hauser et al.[16] used angular brushing to select data that satisfies a limiting angle threshold trend within consecutive axis.

A significant reduction of line cluttering in parallel coordinates is also achieved by data compression. Albazzaz and Wang[17] used independent component analysis to reduce the number of original variables to a small number of statistically independent latent variables. Such a reduction in the number of variables enables the parallel coordinate visualization technique to only demonstrate meaningful data variation. Therefore, a large number of parallel coordinate axes can be reduced to the number of latent variables necessary to provide meaningful data variation.[18]

Latent variables have been effectively used for process monitoring and fault detection in chemical facilities. Process monitoring researchers have made use of latent variables derived from principal component analysis (PCA) to demonstrate the advantages of data compression in fault detection.[19–21] The traditional way to visually monitor a process through data compression techniques is by defining score plots and by plotting the square prediction error (SPE) in time. Score plots tend to only represent two latent variables at a time. Therefore, the number of score plots necessary to represent all possible pairs of l latent variables is given by

Correspondence concerning this article should be addressed to R. Dunia at rdunia@che.utexas.edu.

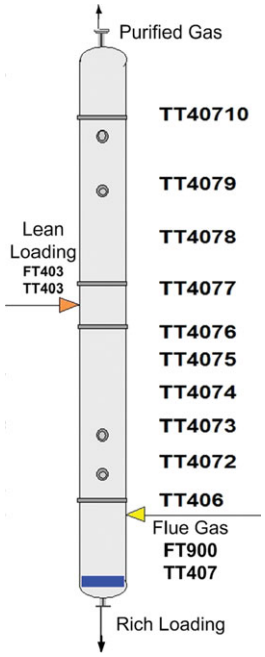

Figure 1. Location of absorber column temperature sensors for multidimensional data.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$\frac{l!}{(l-2)!2!} = \frac{l.(l-1)}{2}.$$

This suggests that score plots cannot be effectively used for data visualization in industrial facilities when more than three latent variables are required to capture the process data variability.

Confidence regions for score plots and SPE have been traditionally determined by the statistical distribution of the latent variables and the residuals.[22–24] Although most of published literature focusses on statistical analysis of past data to calculate the normal operating region of operations, it is important to determine the frequency of false alarms from a particular monitoring system implementation.[25] The visualization of false alarms and their proximity to the actual faults may also provide a clue to predict incipient abnormal operating conditions, especially when dealing with data-driven models.

This article makes use of PCA visualized on parallel coordinates plots. Wang et al.[12] pointed out the difficulty in visualizing more than three principal components in Euclidean plots and suggested the use of parallel coordinates to represent operating regions and established limits of operations for a waste water plant. We have extended the insight given by Wang et al. when using parallel coordinates visualization for process monitoring by analyzing the shape of the confidence region of operations obtained from PCA statistical tests. Furthermore, this work provides case studies where several fault events and models are analyzed simultaneously to have a global perception of the PCA capabilities, not only to detect but also to reduce the number of false alarms.

This article is organized as follows: next section will demonstrate the motivation of using parallel coordinates in chemical process applications. It will also illustrate the way confidence limits are shown when using parallel coordinates as a visualization tool. Dynamic data visualization is also demonstrated in this section when using the same variables lagged by multiples of the data sampling time. The following section introduces the use of principal components inside parallel coordinates visualization, denoted by PC^2. Case studies to predict compressor surge and column flooding using PC^2 are demonstrated. Conclusions and future work are provided in the last section.

Parallel Coordinates for Data Visualization

Motivation

In parallel coordinate visualization,[26–28] a data point consists of N dimensions or coordinates, not necessarily independent. Such a point may represent an operating condition described by sensors, an event outcome defined by several features or a sample characterized by few measurements. Regardless of each coordinate meaning, in parallel coordinates each dimension is drawn as a vertical (or horizontal) line, and a point is visualized as a poly-line that intercepts each axis/coordinate at the appropriate location.

To explain the use of parallel coordinates in chemical processes, consider the eleven temperature measurements along the CO_2 absorber tower depicted in Figure 1. Temperature measurements are available every 5 min for each individual sensor, as illustrated in Figure 2a. This figure provides a time trend temperature at the different absorber locations. However, the trend does not help to visualize the shape of the temperature profile along the column.

Figure 2b illustrates the use of parallel coordinates to demonstrate the temperature profile along the absorber column. Each of the seven poly-lines represents a different time sample with 11 temperature measurements. The time axis is the first parallel coordinate, and the temperature parallel axes are in the same order placed in the absorber column (bottom to top). The high and low values and temperature scaling are made the same for all temperature axes.

The use of parallel coordinates not only assists the visualization of the temperature profiles, but also allows the user to append axis without altering the existing configuration. In the case of the absorber column, it is important to determine the cause of the temperature profile changes in time. For this reason, the flue gas and lean loading feed conditions are incorporated in the way shown in Figure 3. Such a figure includes axes that correspond to the transmitter measurements FT900, TT407, FT403, and TT403 depicted in Figure 1. Such measurements represent the flow rates and temperatures of the flue gas and lean loading flows to the absorber tower. Because the flow rates do not change during the time data have been collected, and because the column temperature profile follows the one for the lean loading temperature, it is concluded that TT403 represents the main cause of the column temperature variation during the period of time the data was collected.

This use of parallel coordinates provides a significant help in searching conclusions based on data visualization. It is important to highlight that the order of the coordinates in this example was chosen based on the sensor type and location. However, in a more general case, the coordinates order might impact the visualization of the results. The next subsection provides the basic tools to define confidence regions and to demonstrate dynamic data profiles in parallel coordinates.

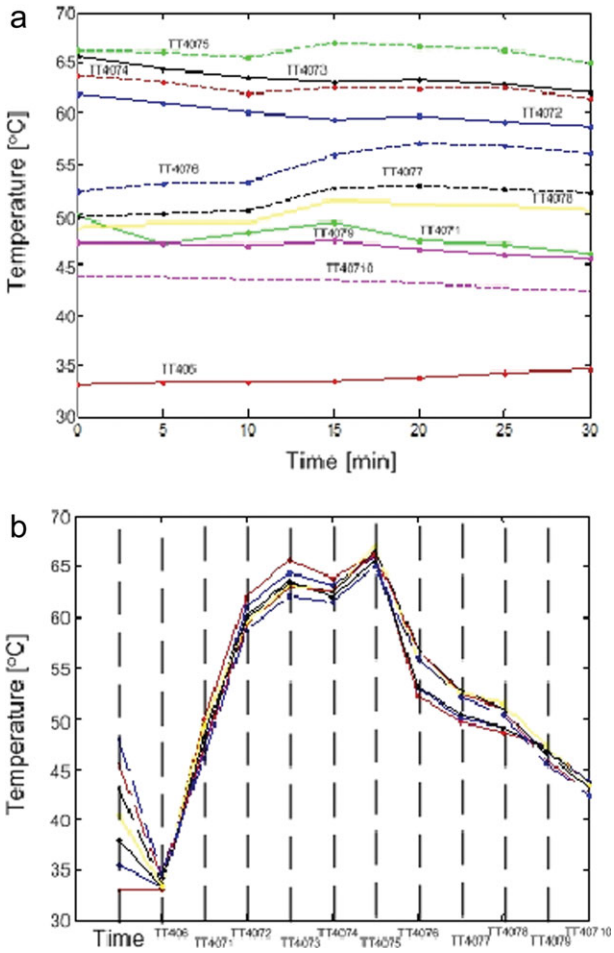

Figure 2. Comparison between the use of time trends with parallel coordinates in the visualization of the temperature profiles along the absorber column.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Confidence regions

Normal process operations are statistically defined by the fulfillment of inequality constraints that involve process variables. The limits used in such inequality constraints are based on statistical analysis of past process data. Confidence regions provide a visual representation of such inequality constraints, and they tend to be multidimensional when more than one variable is involved.

The borderline of a confidence region of operation is determined by the shape that results from the data that satisfy the normal operation inequality constraints. For example, consider the case of two independent variables, x_1 and x_2, with limits l_1 and l_2, respectively

$$\left|\frac{x_1}{l_1}\right| \leq 1 \quad \left|\frac{x_2}{l_2}\right| \leq 1$$

The operating region defined by these two inequality constraints is depicted using Euclidean coordinates and parallel coordinates in Figure 4. Euclidean coordinates (top-left) provides the traditional rectangular area while parallel coordinates (top-right) show a trapezoidal shape. The advantage of the latter approach is the ability to append many trapezoids

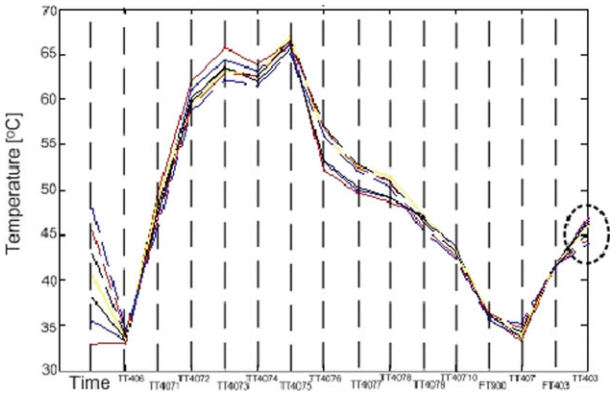

Figure 3. Inclusion of the flue gas and lean loading flow properties with the set of parallel coordinates used in Figure 2b.

The incorporation of such new axes helps to determine the cause of the column temperature profile variation in time. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

and to demonstrate the operating regions for N independent constraint variables, as illustrated in Figure 4 (bottom).

Inequality constraint expressions that involve more than one process variable are potential sources of measurement dependency to be satisfied in order to keep the process under normal operating conditions. To illustrate the effect of having multiple variables in one constraint expression, consider the following set of inequality constraints

$$\left|\frac{x_1 + x_2}{l_1}\right| \leq 1 \quad \left|\frac{x_1 - x_2}{l_2}\right| \leq 1$$

which is equivalent to

$$-l_1 \leq x_1 + x_2 \leq l_1 \quad -l_2 \leq x_1 - x_2 \leq l_2$$

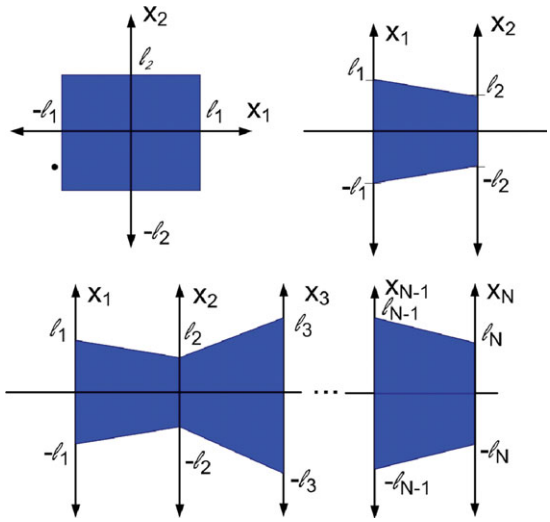

Figure 4. Operating regions defined by independent variable constraints.

Parallel coordinates provides a trapezoid shape (top-right) that can be sequentially repeated $N - 1$ times for N independent variables (bottom). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

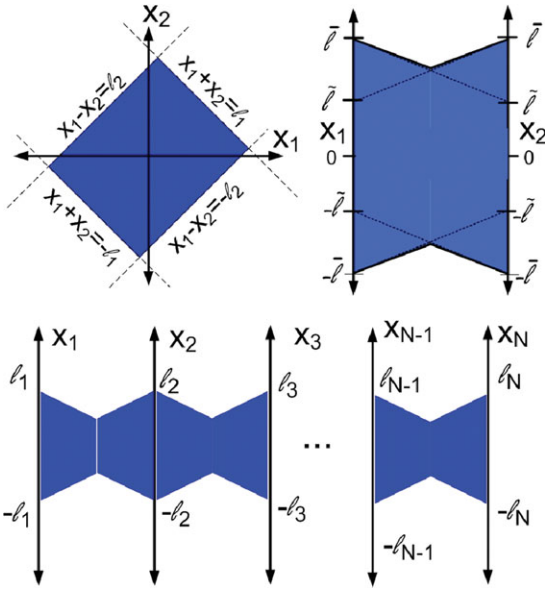

Figure 5. Operating regions defined by two dependent variable constraints. Parallel coordinates provide a bowtie shaped region (top-right) that can be sequentially repeated *N* − 1 times when *N* dependent variables are considered (bottom).

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

These last expressions represent a linear transformation of the univariate constraint case that results into a rotated rectangular region, as illustrated in Figure 5 (top-left). The range calculation of x_1 and x_2 is obtained by the solution of the LP problem in which x_i, where $i \in \{1, 2\}$, is maximized (or minimize) subject to the constraints listed above. The solution gives the pair defined by

$$(\max_{x_j} x_i, x_j) = -(\min_{x_j} x_i, x_j) = (\bar{l}, \tilde{l}), \quad i,j \in \{1,2\}, i \neq j \quad (1)$$

where $\bar{l} = 0.5(l_1 + l_2)$ and $\tilde{l} = 0.5(l_1 - l_2)$. The representation of such a region into parallel coordinates provides "bowtie shaped" confidence regions, as it is shown in Figure 5 (top-right).

A practical way to distinguish the variables based on their particular variance is by assigning the limit l_i to its corresponding variable x_i in the following manner

$$|\frac{x_1}{l_1} + \frac{x_2}{l_2}| \leq 1 \quad |\frac{x_1}{l_1} - \frac{x_2}{l_2}| \leq 1$$

which results in the following constraints to the LP

$$-1 \leq \frac{x_1}{l_1} + \frac{x_2}{l_2} \leq 1 \quad -1 \leq \frac{x_1}{l_1} - \frac{x_2}{l_2} \leq 1$$

The solutions when solving the LP that maximizes (or minimizes) x_i based on these constraints give

$$(\max_{x_j} x_i, x_j) = -(\min_{x_j} x_i, x_j) = (l_i, 0), \quad i,j \in \{1,2\}, i \neq j \quad (2)$$

Notice from Figure 6 (left) that the bowtie shaped confidence region for this solution conserves its horizontal symmetry axis, but not the vertical one due to the disparity between the two limits. In a similar manner, if N linearly independent multivariable inequality constraints of the form

$$\left|\sum_{i=1}^{N} \alpha_{ij} x_i\right| \leq 1 \quad \Rightarrow \quad -1 \leq \sum_{i=1}^{N} \alpha_{ij} x_i \leq 1 \quad \text{for} \quad j = 1, \ldots N$$

are used to define a closed operational confidence region, then $N - 1$ sequential 'bowtie shaped' regions are expected to show similar confidence region of operation in parallel coordinates, as depicted in Figure 5 (bottom). The coefficients α_{ij} inherit the role of the reciprocal limits $1/l_j$ and are made to define a closed-feasible region of operation. Nevertheless, the general form for the N linearly independent constraints based on α_{ij} coefficients may provide conditions for variables that are completely independent of the rest. As an example, if the variable x_i is independent of the other ones, i.e., $\alpha_{ij} = 0 \ \forall i \neq j$, then "trapezoidal shaped" confidence regions are formed around such a variable in the parallel coordinate plot.

Ellipsoids are the most popular way to define convex confidence regions when applying statistical data analysis to define normal operating conditions in Euclidean coordinates. A two-dimensional (2-D) ellipsoid is defined by the following expression

$$\left(\frac{x_1}{l_1}\right)^2 + \left(\frac{x_2}{l_2}\right)^2 \leq 1 \quad (3)$$

The use of ellipsoidal confidence regions in Euclidean coordinates results in "nozzle shaped" patterns in parallel coordinates, as is illustrated in Figure 6 (right). This nozzle pattern has the same limits than the ones obtained for bowtie shaped regions but with rounded borders. The limits are identical because the "max" and "min" solutions are the same in both cases. Figure 6 demonstrates the use of same parameters l_1 and l_2 for the bowtie (left) and nozzle (right) shapes. The figure also shows that the latter has a larger confidence region than the former. The difference between these two region edges is due to the nonlinearity nature of the ellipsoid constraint. An analytical expression of such a difference is given by the line that connects the points from coordinate x_1 to x_2 in parallel coordinates

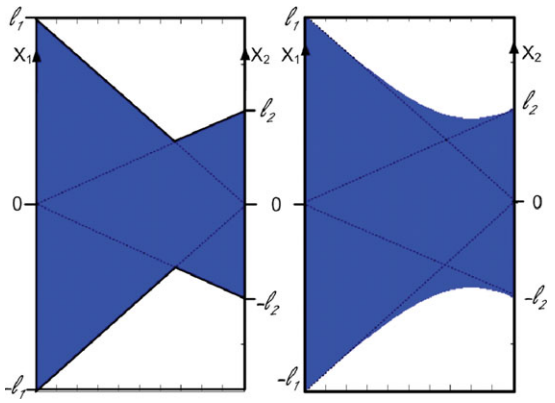

Figure 6. Comparison between the bowtie shaped (left) and nozzle shaped (right) confidence regions when using parallel coordinates.

Both graphs share the same statistical parameters, l_1 and l_2. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

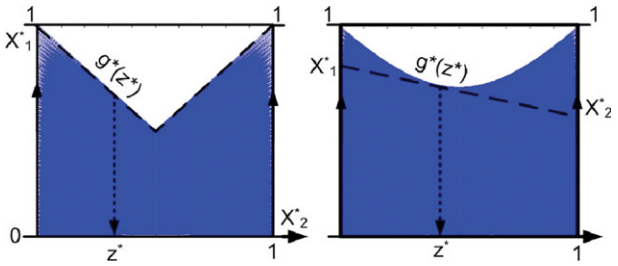

Figure 7. Calculation of the bowtie shaped (left) and ellipsoidal (right) confidence regions when using parallel coordinates.

The same linear objective function $g^*(z^*)$ is maximized subject to different type of constraints. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

$$g(z) = \left(\frac{x_2 - x_1}{\Delta}\right) z + x_1 \tag{4}$$

where z and Δ represent the horizontal axis and the distance between the parallel coordinates, respectively. For convenience we define the dimensionless form of $g(z)$, denoted by $g^*(z^*)$,

$$g^*(z^*) = x_2^* z^* + x_1^*(1 - z^*) \tag{5}$$

where $z^* = \frac{z}{\Delta}$, $x_1^* = \frac{x_1}{l_1}$ and $x_2^* = \frac{x_2}{l_2}$. The optimization objective function to maximize (or minimize) to determine the analytical expression of the confidence region in parallel coordinates is posted in the following manner

$$\max_{x_1^*, x_2^*} g^*(z^*) \tag{6}$$

where $0 \le z^* \le 1$. In the case of rectangular confidence regions, the linear constraints for the optimization problem are given by

$$-1 \le x_1^* + x_2^* \le 1 \quad -1 \le x_1^* - x_2^* \le 1$$

while for ellipsoidal confidence regions

$$(x_1^*)^2 + (x_2^*)^2 \le 1$$

The optimal solutions for these two type of constraints can be visualized in Figure 7. The solution goes along the linear constraints for the rectangular confidence region, i.e., bowtie-shape region in parallel coordinates

$$g^*(z^*) = \begin{cases} 1 - z^* & 0 \le z^* \le 0.5 \\ z^* & 0.5 < z^* \le 1 \end{cases} \tag{7}$$

while it represents a nonlinear function for the ellipsoidal confidence region, i.e., nozzle-shape region in parallel coordinates

$$g^*(z^*) = \sqrt{(z^*)^2 + (1 - z^*)^2}, \quad \text{for } 0 \le z^* \le 1 \tag{8}$$

The extension of this last solution to N parallel coordinates is obtained by solving $N - 1$ independent optimization problems of the form

$$\max_{x_j^*, x_{j+1}^*} g_j^*(z^*), \quad \text{for } 1 \le j \le N - 1 \tag{9}$$

where

$$g_j^*(z^*) = x_{j+1}^* z^* + x_j^*(1 - z^*), \quad \text{for } 0 \le z^* \le 1 \tag{10}$$

subject to

$$\sum_{i=1}^{N} x_i^{*2} = 1 \tag{11}$$

where $x_j^* \equiv \frac{x_j}{l_j}$ and $z^* \equiv \frac{z}{\Delta}$. Because we are only interested in the confidence region g_j^* defined between the pair of coordinates x_j^* and x_{j+1}^*, the solution to this problem is identical to the case of using two variables at a time, while the remaining variables are set to zero, i.e.

$$g_j^*(z^*) = \sqrt{(z^*)^2 + (1 - z^*)^2}, \quad \forall j \in [1, N-1] \tag{12}$$

Figure 8 (top) illustrates the solution of the optimal region g_j^*, whereas Figure 8 (bottom) demonstrates the superposed solution for all consecutive pairs of parallel coordinates. Although this type of confidence region does not take into account the probabilistic distribution of the data, nor the multivariable dependency of variables that are not adjacent in the plot, it represents an interesting principle for parallel coordinate data visualization.

The three classes of regions presented above are summarized in Figure 9. In Euclidean coordinates (left) the confidence regions represented by the rectangular, ellipsoid and rotated-rectangular shapes corresponds to the yellow, blue, and red colors, respectively. Their equivalent representation in parallel coordinates are the trapezoid, nozzle, and bowtie shape, as it is illustrated in Figure 9 (right). Notice the transformation of the diagonal points a, b, and c in Figure 9 (left), to the horizontal lines on the parallel coordinate representation (right).

The use of polar coordinates can provide some advantages in the specification of ellipsoidal confidence regions. Two parallel coordinates that correspond to the radius r and angle

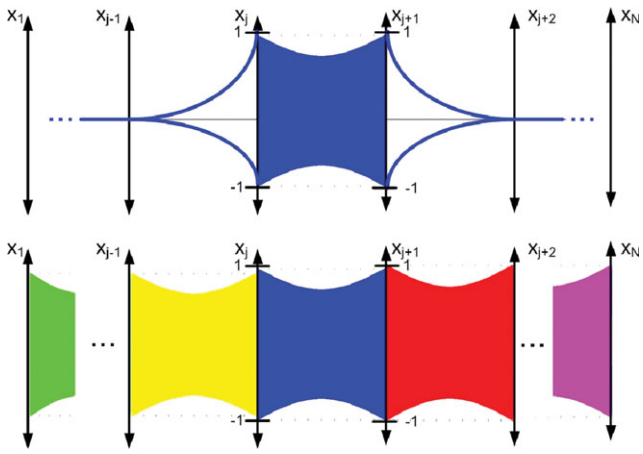

Figure 8. Solution to the parallel coordinate confidence region.

The nozzle-shape obtained from a pair of consecutive coordinate solution (top) is superposed to extend this pattern along all $N - 1$ segments (bottom). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

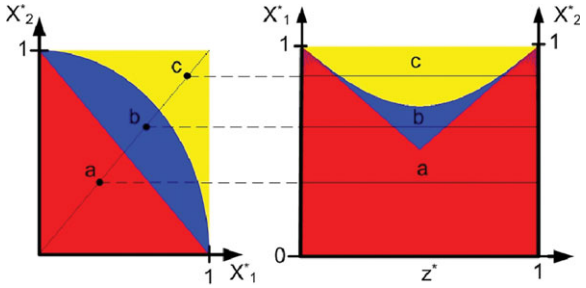

Figure 9. Euclidean (left) and parallel coordinate (right) representation of the three type of confidence regions.

Diagonal points a, b, and c in Euclidean coordinates are transformed into horizontal lines in parallel coordinates. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

θ are used to specify points in polar coordinates. The use of polar coordinates permits depicting points within a threshold from the confidence region limit, as illustrated in Figure 10. In this figure, points within the radius range $R - \delta \leq r \leq R$ are shaded in polar coordinates (left) and parallel coordinates (right). Such shaded regions can represent points that are within δ radial units from the confidence region edge.

Next the use of parallel coordinates for dynamic sampled data is demonstrated.

Discrete dynamic data

Parallel coordinates can be arranged to demonstrate not only the current outcome of a sensor measurement, but also the measurements obtained from previous sample times. In this way, it provides a profile similar to time dependent plots where only a few past samples are taken into account. As an example, consider the following single input-single output dynamic system

$$y(k) = ay(k-1) + bu(k-1) + e(k) \tag{13}$$

where u and y represent the input and output of the system. The scalar e denotes the noise and k is the sample index. The parameters $a \in (-1, 1)$ and b are considered constant. Figure 11 demonstrates the response for a unit step input in u. The solid dots of the same color represent the response at the same

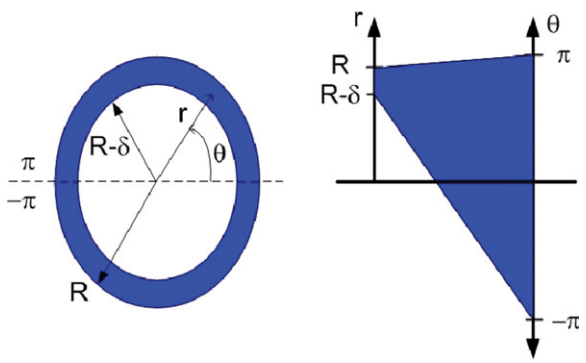

Figure 10. Confidence region represented in polar and parallel coordinates.

The shaded area highlights the portion of normal data within δ radial units from the region limit, R. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

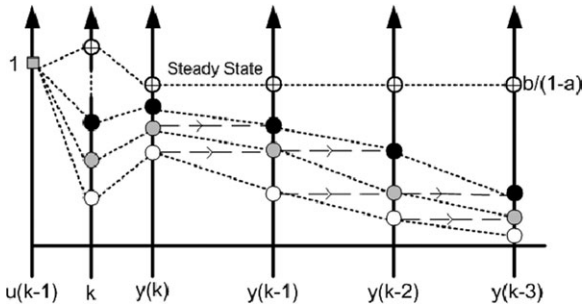

Figure 11. Dynamic sample data represented in parallel coordinates.

Solid points represent the deterministic portion of the response given by Eq. 13.

time stamp k. As new samples are collected, the plot updates accordingly by moving the data points towards the following axis on the right.

For a step input change, the dynamic response for all parallel coordinate axes that represent the output y will end at the steady state line, represented by the symbol points $\oplus$ in Figure 11. The steady state can be also calculated by solving for y in Eq. 13.

Principal Components in Parallel Coordinates

Principal components have been extensively used for process monitoring and fault detection in industry.[29] One of their limitations in process applications is the visualization of score plots in 2-D graphs. Therefore, the use of l principal components requires $\frac{l(l-1)}{2}$ 2-D plots to depict operating points outside normal operating conditions. Furthermore, the inspection of all possible 2-D score plots does not provide a full screening of points outside the confidence region. As an example, consider the use of three principal components and the Hotelling T^2 confidence region,[30,31] given by

$$\left(\frac{t_1}{l_1}\right)^2 + \left(\frac{t_2}{l_2}\right)^2 + \left(\frac{t_3}{l_3}\right)^2 \leq 1 \tag{14}$$

A point $\mathbf{x} \in \Re^N$ with scores parameterized by $t_i = \alpha l_i$ verifies the Hotelling T^2 condition if $\alpha^2 \leq \frac{1}{3}$. However, such a condition is reduced to $\alpha^2 \leq \frac{1}{2}$ when projecting into the 2-D score plots. Notice that projecting into 2-D score plots is equivalent to setting $t_j = 0$, where j denotes the score index not considered in the projection plane. Therefore, a point $\mathbf{x} = \alpha\mathbf{t}$, where $\mathbf{t} = [t_1 \quad t_2 \quad t_3]^T$ and $\frac{1}{\sqrt{3}} \leq \alpha \leq \frac{1}{\sqrt{2}}$ results in an abnormal operating point that cannot be visualized by any of 2-D score plots. Furthermore, the greater the number of principal components, the less likely abnormal operating conditions can be visualized using 2-D score plots.

Parallel coordinate graphs provide the required plotting features needed for process monitoring using principal components to simultaneously visualize all scores in one plot. The use of as many parallel coordinates as principal components permits one to represent an nth dimensional point using l parallel coordinates, each representing a score or latent variable, as illustrated in Figure 12. Besides the scores, the Hotelling T^2 and the SPE, defined by

$$\mathrm{SPE}(k) \equiv \|\mathbf{x}(k) - \mathbf{P}\mathbf{t}(k)\| \tag{15}$$

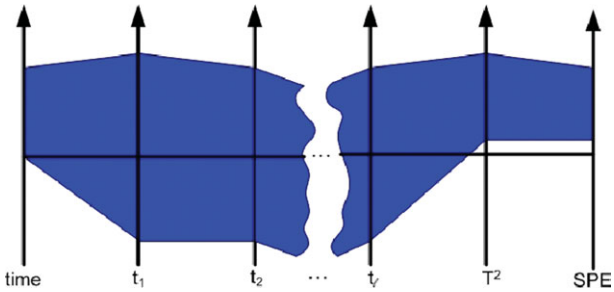

Figure 12. Principal component analysis in parallel coordinates.

All scores and statistical tests can be visualized simultaneously in this type of plots. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

and

$$T^2(k) \equiv \sum_{i=1}^{l} \left(\frac{t_i(k)}{l_i} \right)^2 \tag{16}$$

can also be incorporated as parallel coordinates. In the expression above, the columns of the matrix **P** represent the principal directions on which most of the data variation takes place.[18] Figure 12 shows how these statistical indexes can be made part of the parallel coordinates plot. Furthermore, the tests can consider models with a different number of principal components to determine the most suitable value for l to detect a fault. Such a selection can be based on how well faults are detected and how many false alarms are flagged for different values of l.

Another important feature available when using parallel coordinates for process monitoring is the possibility of highlighting data across all coordinates when a particular condition is satisfied. In the specific case of fault detection and isolation using principal component and statistical tests, the process engineer can determine under which conditions such tests are not verified and inspect the contribution of each score or residual for fault isolation. Figure 13 illustrates the use of this feature by coloring the data region that violates the Hotelling T^2 with gray. Therefore, all points with $T^2 > 1$ are selected along the T coordinate axis and highlighted in gray across all parallel coordinates. It is important to notice that the first score t_1 shows as the one with the most significant contribution to this fault. Therefore, the elements of the first principal component can provide a clue for the identification of such a fault.[32]

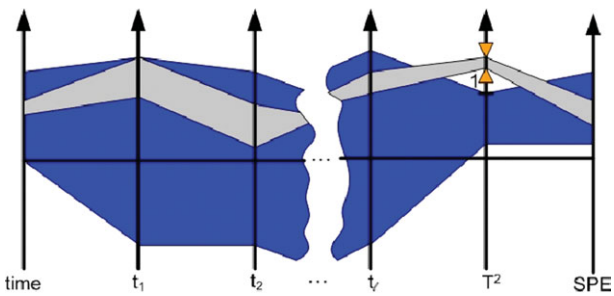

Figure 13. Use of principal component with parallel coordinates for fault detection and identification.

The points for which $T^2 > 1$ are highlighted in gray to determine how this abnormal condition propagates across all coordinates. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

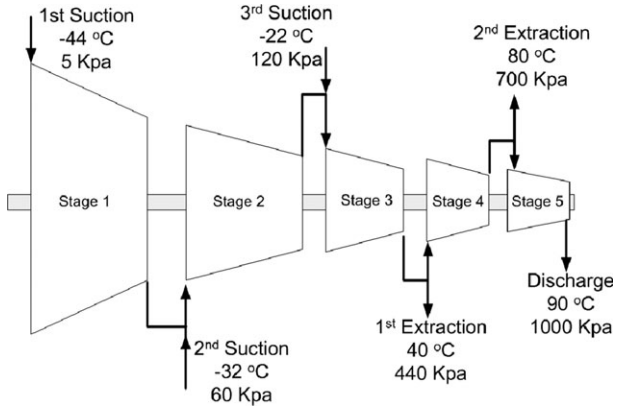

Figure 14. Compressor used for propylene refrigeration system.

It consists of five compression stages with a total of three suctions, two extractions and one discharge stream.

Case Studies

Compressor surge in propylene refrigeration systems

Closed loop propylene refrigeration systems are used to chill process streams to temperatures below −40°C in olefin plants, as illustrated in Figure 14. The compressor provides four refrigeration stages to separate and produce ethylene and propylene at polymerization grade. The compressor shown in Figure 14 has three suctions, two extractions and one final discharge.

Compressor surge is characterized by dynamic unstable operations due to low volumetric internal flows for the amount of power provided to the compressor. In a multistage compressor, there are rotor blades attached to the moving axis that provide momentum to the fluid, and stator blades that transform such a momentum into pressure increase. Because the compressor increases pressure in the flow direction, low internal flows can reverse within compression stages, causing surge. Under surge conditions vibrations can exceed the maximum operating limits and the equipment overheats, which causes pulsing discharge flows. Although this phenomena should be avoided at all cost, optimal operations are close to surge conditions.

The antisurge system consists of opening a recycle line to increase the flow rate inside the compressor. It represents a safe but very inefficient mode of operation, as all compressed flow is decompressed and sent back to the compressor suction lines. This costly operational adjustment and the difficulty to predict surge by monitoring the flows around the compressor create the need for an advanced process monitoring technique.

Figure 14 demonstrates that the intermediate suction and extraction flows do not provide the total flow across the compressor stages. The lack of measuring the main physical condition that causes surge indicates that this phenomena is not easy to predict. Figure 15 provides the pressure, flow and temperature of the suction and extraction lines one day before the surge event. Although surge is about to occur at time zero, there is no major disruption in the stream data

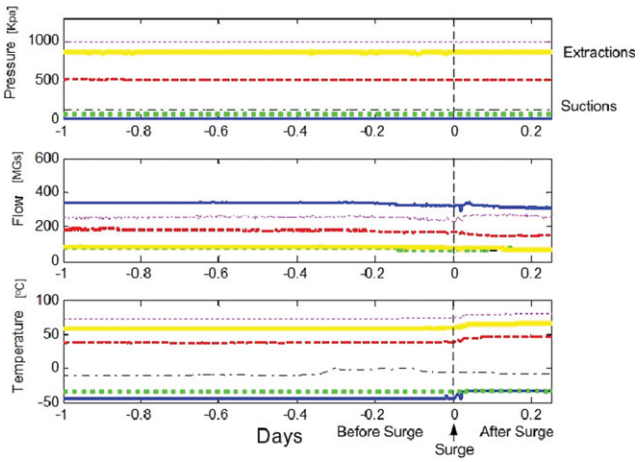

Figure 15. Compressor suction and extraction stream measurements (pressure, flow rate, and temperature profiles).

Time zero represents the instant at which the compressor starts to surge. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

trends around the compressor. Therefore, surge cannot be predicted by simply using individual stream measurements around the compressor.

A detailed analysis of the measurement trends close to surge conditions demonstrates that measurement oscillations tend to be more pronounced as surge time approaches. Therefore, signal characteristics that determine amplitude and frequency of oscillations were included together with pressure, flow rate and temperature measurements in the development of an empirical multivariable model for process monitoring using PCA. A total of 25 surge events were considered in the development of such a model. All data collected previous to one day before each surge event account as part of the compressor normal operations. Abnormal data features that lead to surge early detection are withdrawn from the measurements collected 2 hours before the compressor is in surge.

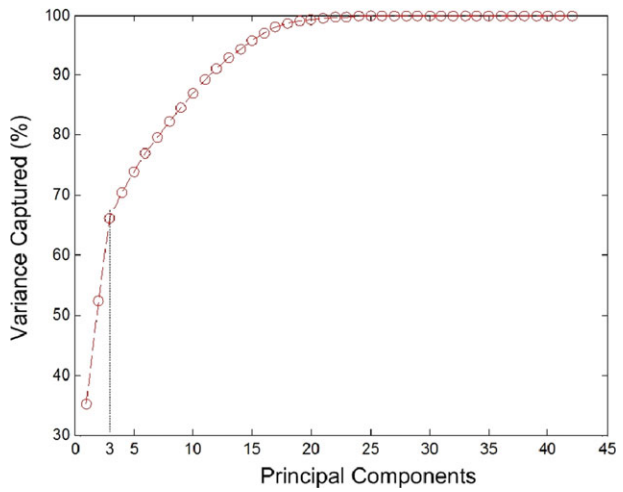

Figure 16. Variance captured by different number of principal components in the model.

A total of 42 variables are used and 70% of the data variance is captured with only three principal components. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

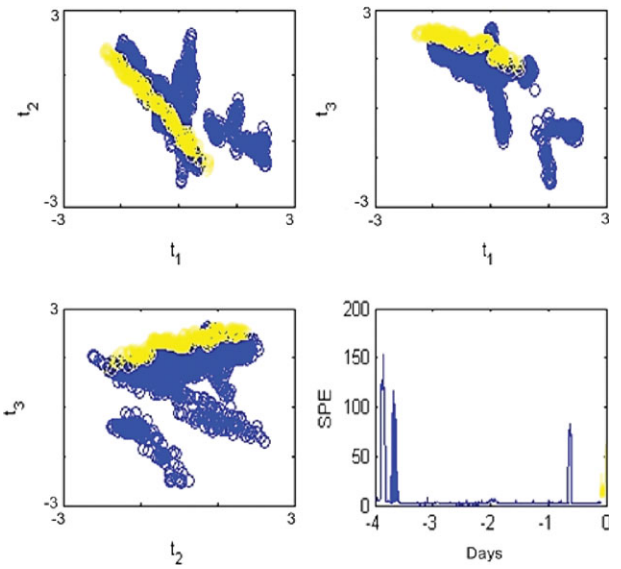

Figure 17. Score and SPE plots for the 12 surge event.

The blue points in the score plots represent data collected between four and one day before surge, while yellow points correspond to data collected within 2 hours before the surge event. Data collected at least one day before the surge event is considered normal, while the one taken within 2 hours from surge may provide an insight of this deleterious event to come. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Figure 16 demonstrates the amount of variance captured by the PCA model for normal operating data when using a different number of principal components. Notice that a model with three principal components captures close to 70% of the data variance of 42 variables, and the contribution provided by subsequent principal components is negligible compared to these first three components. Therefore $l = 3$ and three 2-D score plots will be necessary to visualize abnormal operating conditions for the Hotelling T^2 test.

The three 2-D score plots and the SPE for the 12 surge events are depicted in Figure 17. The blue points in the score plots represent data collected between four and one day before surge, while yellow points correspond to data collected within 2 hours before the surge event. Notice that t_3 and t_1 are relatively large and small, respectively, within 2 hours from surge in comparison with the data collected more than one day before surge. The SPE plot shown in Figure 17 (right-bottom) does not provide major clues regarding the surge event at time zero.

Although the score plots have been the standard way to visualize process monitoring and abnormal operating conditions using PCA, the outcome is not clear in applications where the potential fault should be predicted in advance. Furthermore, process engineers are required to observe several 2-D plots simultaneously to detect faulty conditions.

Figure 18 illustrates the use of principal components in parallel coordinates (PC^2) for all 25 surge events. There is a total of 15 parallel coordinates that correspond to time before surge, surge event, scores, Hotelling T^2 and SPE from two to five principal components. Such a plot includes not only all surge events at once, but also provides the results for models with different number of principal components. Portions of the data can be conditionally highlighted to eliminate the cluttering of lines when some specifications are

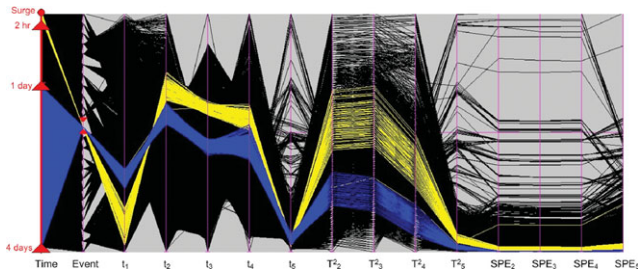

Figure 18. Use of principal components with parallel coordinates for 25 surge events.

The parallel coordinates considered here include scores, T^2 and SPE tests for two to five principal components. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

verified. In the case of Figure 18, data within 1 hour from the 13th surge event has been highlighted in yellow. For the same surge event, data recorded from four to one day before surge was highlighted in blue.

The results illustrated in Figure 18 show that T_3^2, which corresponds to the Hotelling T^2 test for three principal components, increases significantly from the time range of four to one days before surge (blue) to 2 hours within surge time (yellow). Therefore, such an index could have been used to predict the coming of surge 2 hours before the 12th surge event. Furthermore, Figure 18 is helpful to point out that the first and third scores (t_1 and t_3) are crucial in early surge detection. This is because t_1 is consistently small and t_3 is consistently large within 2 hours from surge than one to four days from surge time. Finally, models that make use of two, four, and five scores, as well as the SPEs are shown in Figure 18 as ineffective in the early detection of this phenomena.

In summary, the results demonstrated in a parallel coordinate plot provides useful process monitoring information regarding all surge events within days before surge occurs and for models using different number of principal components. Such information consist of Hotelling T^2, SPE and scores contributions, useful for early detection and identification of a surge event. The next case study illustrates the use of parallel coordinates in an atmospheric distillation column to predict flooding.

Distillation column flooding

Distillation column flooding occurs when liquid coming from upper sections of the column fills up the interfacial liquid–vapor spaces of the column packing. The phenomena can happen at any level of the column due to large drag forces of the vapor coming upward on the falling liquid. The drag forces oppose the force of gravity and slow the falling liquid flow to the stagnant point, called the flooding point. Liquid-vapor equilibrium is significantly deteriorated along the distillation column due to flooding, and product purity deviates from product specifications. However, plant operations are most efficient near flooding conditions. Therefore, while column operators intend to maximize throughput, adjustments in operating conditions can result in flooding.[33]

Figure 19 illustrates the distillation column with some of the nominal measurement values for operation and monitoring. It consists of more than 30 equilibrium stages with a gas heater reboiler and two overhead condensers (not included in the figure). Given the high fluid viscosity at low temperatures, flooding can happen at any equilibrium stage including the upper section of the column, where temperatures tend to be lower compared to the rest of the column.

Process data were collected every minute during 4 months. There was a total of eight flooding periods approximated to round days because it is not easy to determine either the exact starting or final time, nor the location of flooding. Process variables with significant variation were considered for model development and normalized according to their average and standard deviation. There were 43 sensor measurements considered for process monitoring, which include pressures and temperatures along the column and feed/product flow meters.

Figure 20 demonstrates the use of parallel coordinates to visualize all 43 measurements along the 4-month operating period. The eight flooding event data sets are highlighted in yellow and represent 14% of the data collected. There is a significant number of days where the column operated under flooding conditions, to the point that a flooding model was later developed to determine if the column is at such abnormal operating condition. Therefore, a PCA model based on faulty data will detect a potential faulty condition when the SPE or the Hotelling T^2 index are small.

It is important to highlight from Figure 20 that flooding data seem to span throughout all variable ranges. Therefore, there is no unique physical variable that can be used to detect flooding for all eight events. Such an outcome suggests the use of multivariable process monitoring methods, such as PCA. Figure 21 illustrates the amount of data variance captured by the model based on the number of principal components. The data considered in this model consist of seven out of eight flooding events. The fourth event was left from the PCA model design to determine fault detectability without using all the faulty data available during the 4- month period. Therefore, the fourth event is used to validate the fault detection technique proposed in this work.

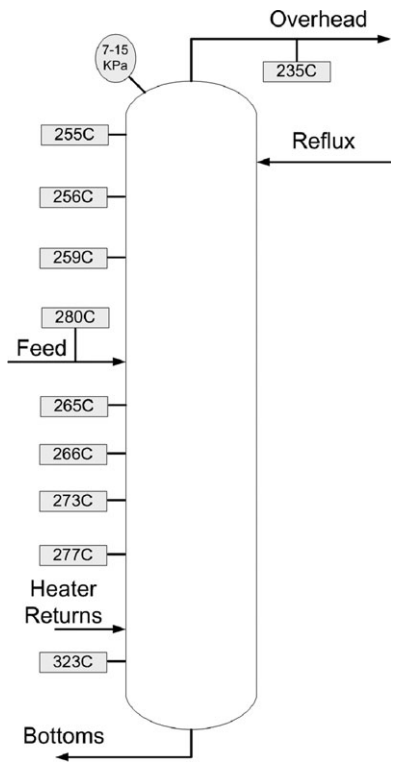

Figure 19. Distillation column schematic.

Flooding can occur at any section of the column due to high fluid viscosity.

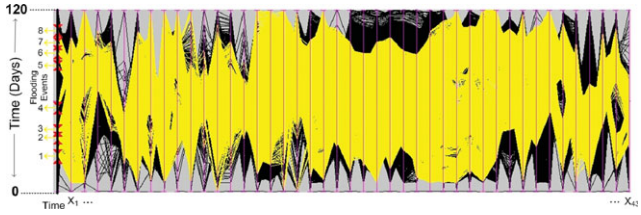

Figure 20. Use of parallel coordinates for flooding column data visualization.

A total of 43 variables are considered, and the yellow regions represent flooding periods. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The results shown in Figure 21 demonstrates that four principal components can capture close to 80% of the data variance during seven flooding events. The use of $l = 4$ indicates a total of six score plots that may not be helpful to visualize conditions in which the column is not flooding. This is because most of normal operating points are expected to have large scores that may not be tractable on zero centered score plots. For this and other reasons pointed out in this article, principal components with parallel coordinates are used for process monitoring.

Figure 22 demonstrates the use of principal components in parallel coordinates. The seven parallel axes used in the figure consists of time, four latent variables or scores, the Hotelling T^2 index and the square prediction error, SPE. The yellow portion of the plot corresponds to the flooding data points with SPE $\leq$ 5. Notice that all flooding periods, including the fourth flooding period which was not considered in the development of the model, contains data that satisfy the condition of having SPE $\leq$ 5. This result indicates that the model is capable of detecting flooding when SPE $\leq$ 5 for faulty data not seen during model development.

Figure 23 highlights the fourth flooding event, which was not used for the PCA model development. The portion highlighted in yellow shows the fourth flooding event data that verifies SPE $\leq$ 5, while blue indicates data with SPE $>$ 5.

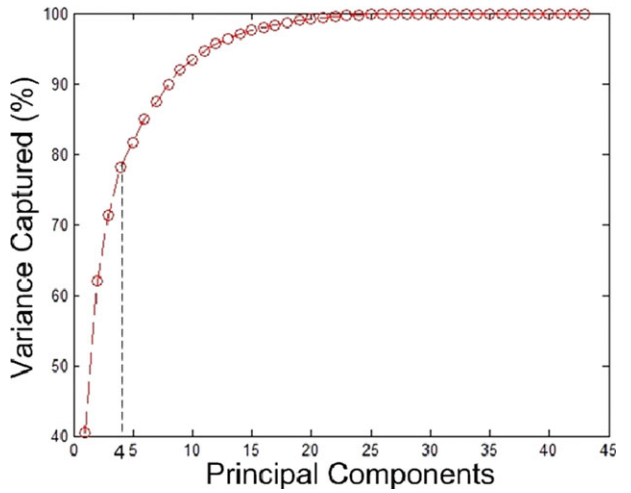

Figure 21. Data Variance captured by the model for different number of principal components.

Four principal components were selected for the column flooding data to provide a total of 80% variance captured by the model. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

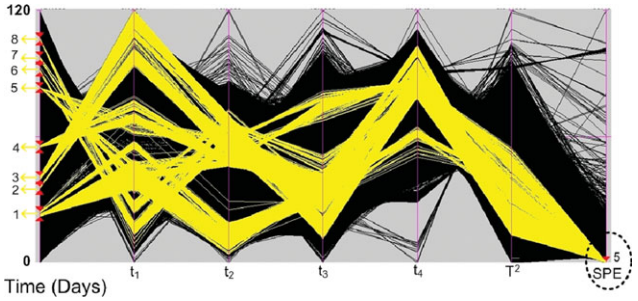

Figure 22. Flooding data visualization using principal components in parallel coordinates, PC^2.

All flooding data with SPE $\leq$ 5 are highlighted in yellow, while the remaining portions of data are considered part of normal operations. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Note that t_4 tends to be positive for the fourth flooding period, and that a significant number of points in this period have a low SPE but not necessarily below five (blue data). This type of conclusions can be withdrawn when the SPE and scores are visualized in the same plot, as it is the case of PC^2.

Figure 24 represents details of the fourth flooding period to determine when was flooding detected. This detailed figure demonstrates that the fault is detected at the time the data lines turned yellow (SPE $\leq$ 5), around t = 55.8 days. This time corresponds to 0.8 days after the flooding period starts. It is important to mention that the flooding periods were given in round day numbers and flooding might have been initiated at the precise instant when SPE jumps below five. Nevertheless, this work is intended to not only show the capabilities of PCA but also the bright visualization feature of PC^2 when compared to the traditional use of score and SPE plots.

The elimination of false alarms represents a very important task for reliable process monitoring methods. An excessive number of false alarms may force the operators to turn the flooding detection system off permanently. Parallel coordinates provide a powerful way to illustrate the frequency of false alarms by highlighting all data points outside the flooding periods that satisfy the condition SPE $\leq$ 5. Figure 25

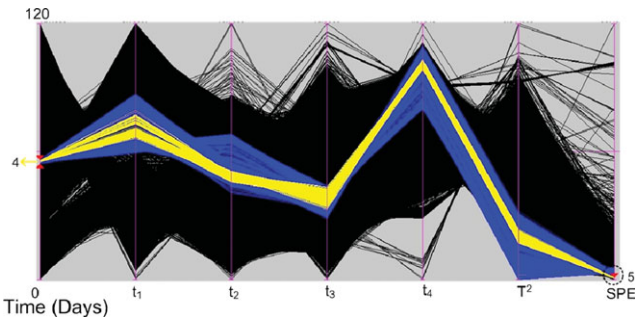

Figure 23. Fourth flooding event highlighted in yellow and blue for data points with SPE below and above five, respectively.

PC^2 provides an overall view of the scores, T^2 and SPE for any portion of data in study. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

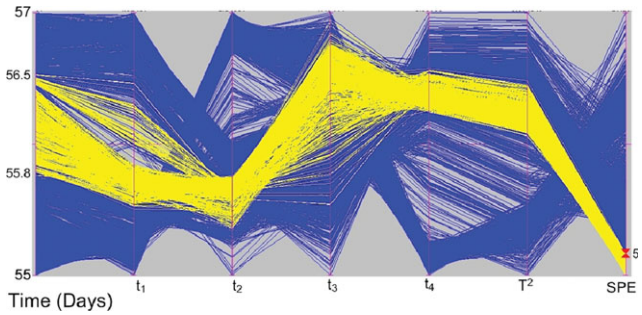

Figure 24. Fourth flooding period detail in yellow and blue for data points with SPE below and above five, respectively.

This plot suggests that the fourth flooding detection spans from 55.8 to 56.5 in day units. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

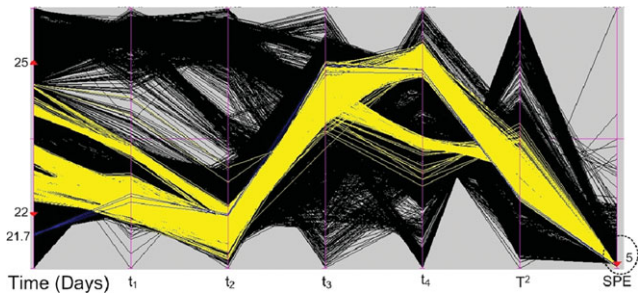

Figure 26. First flooding event detail.

Flooding period was expected between days 22 and 25 of recorded data. A false alarm is detected at $t = 21.7$, which well could be a prediction of the first flooding event happening several hours later. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

illustrates in blue the false alarms, i.e., data points outside the flooding periods with small SPE.

The number of false alarm points is about 0.3% of the total amount of data recorded. Although this number seems to be high, false alarm points tend to occur in consecutive samples. Figure 25 demonstrates six false alarm periods for the 4 months data. Therefore, from seven flooding alarms detected by the SPE, three of them might be false. Nevertheless, such false alarms tend to happen close to the flooding periods, which suggests that the false alarm might well be a prediction or a remanent of a flooding event. This false alarm analysis can only be done off-line because flooding periods have been predetermined. Nevertheless, the conclusions withdrawn from this analysis can be well implemented during on-line application of this technique.

Figure 26 is used to highlight points inside (yellow) and in the vicinity (blue) of the first flooding event that satisfies SPE ≤ 5. It is clear that the blue lines may well represent an early prediction of the faulty event instead of a false alarm. This can benefit the use of PC^2 in a way that flooding could be avoided by taken preventive action. Further research is needed to determine if flooding could be predicted and avoided by making the necessary operation adjustments in the process.

Conclusions

This work demonstrates the use of parallel coordinates for process monitoring. In particular, the proposed technique called PC^2 consists of the visualization of principal components with parallel coordinates. The insight provided by PC^2 in this research work complements the one proposed by Wang et al.,[12] in which parallel coordinates with principal components were implemented in a waste water facility. This article emphasizes the following features of PC^2:

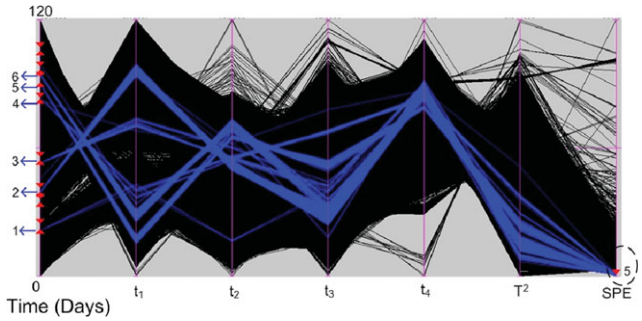

Figure 25. False alarms are represented in blue and consists of data points outside the flooding periods with SPE $\leq$ 5.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

The simultaneous visualization of several faulty events of the same nature

Parallel coordinates permit highlighting common features from different faulty events in the same plot to determine similarities that can point to key data features for fault detection and isolation.

Model comparison when using different number of principal components

Because parallel coordinates permit the inclusion of as many axes as needed, models with a different number of principal components can be compared to determine the most effective way to detect a fault. Therefore, models can be targeted to detect particular faults.

Accurate visualization of points outside the confidence region

The fact that the scores, Hotelling T^2, and SPE are represented in the same plot permits analyzing the different causes of a fault using the same graph. The use of 2-D score plots are very limited as was pointed out in previous publications.[12]

Representation of typical process monitoring confidence regions using parallel coordinates

Confidence region shapes have been characterized based on the type of shapes obtained within consecutive parallel axis. Therefore, this work demonstrates that rectangular and ellipsoidal confidence regions in Euclidean coordinates results in well defined shapes for parallel coordinate representations.

Visualization of data transients using parallel coordinates

Because a process data historian does not necessarily represent steady state conditions, lag measurements can be used as parallel axes to appreciate the dynamic response of a process variable to a manipulated variable change.

Determination of false alarms frequency

The highlighting of false alarms for all fault events permits estimating the frequency and proximity of false

positives to actual process faults, such as the cases of column flooding and compressor surge events. Some false alarms can be considered precursors of a fault.

The demonstration of such significant improvements in process monitoring visualization demonstrates the great potential that parallel coordinates can provide to industrial applications. This user-friendly graphical environment was illustrated in two industrial case studies with actual process data. The first case study demonstrated the use of parallel coordinates with principal components to early detect compressor surge. This application required the comparison of several surge events simultaneously to identify common features among the different fault events. The second case study illustrated how data can be sorted accordingly based on column flood events. In particular, false alarms were highlighted to determine their frequency and proximity to the different flooding events.

The efforts in the development of PC^2 for commercial applications will be focussed on user friendly graphical interface for on-line fault detection and identification. The automated implementation of this technique may represent a challenge to its success as the methodology involves the visualization of data, ordering and arrangement of the parallel coordinates. Another important component to make this tool more successful is the analysis and study of scores in conjunction with the elements of the principal component model matrix to discern particular characteristics of abnormal operating conditions. Such characteristics will guide fault isolation as sampled data is acquired by the distributed control system.

Acknowledgments

The authors gratefully acknowledge the support of Emerson Process Management and the Center for Operator Performance (COP) for this research work. Process data and plant operations insight provided by Peter Vermeer from Suncor and Michael Bell were of great support to test the proposed technique with practical case studies. Finally, this work would have not been possible without the assistance of Dr. Alan Mahoney from Process Plant Computing Limited, who supplied the training necessary for parallel coordinate plot analysis and data management.

Literature Cited

1. Inselberg A, Dimsdale B. Parallel coordinates—a tool for visualizing multidimensional geometry. Presented at the *1st IEEE Conference on Visualization (Visualization 90)*, San Francisco, CA, Oct 23–26, 1990:361–378.
2. Inselberg A. Visual data mining with parallel coordinates. *Comput Stat*. 1998;13:47–63.
3. Dubska M, Herout A, Havel J. PC lines—line detection using parallel coordinates. Presented at the IEEE Conference on Computer Vision and Pattern Recognition, (CVPR), Colorado Springs, CO, Jun 20–25, 2011:1489–1494.
4. Tao W, Lian-Dong Y. Design of the data processing system for parallel dual-joint coordinate measuring machine. Presented at the 4th International Seminar on Modern Cutting and Measurement Engineering, Beijing, Peoples Republic of China, Dec 10–12, 2010.
5. Alsakran J, Zhao Y, Zhao X. Tile-based parallel coordinates and its application in financial visualization. Presented at the Conference on Visualization and Data Analysis 2010, Vol. 7530, San Jose, CA, Jan 18–19, 2010.
6. Steed CA, Fitzpatrick PJ, Jankun-Kelly TJ, Yancey AN, Swan JE II. An interactive parallel coordinates technique applied to a tropical cyclone climate analysis. *Comput Geosci*. 2009;35:1529–1539.
7. Steed CA, Fitzpatrick PJ, Swan JE II, Jankun-Kelly TJ. Tropical cyclone trend analysis using enhanced parallel coordinates and statistical analytics. *Cart Geograph Inform Sci*. 2009;36:251–265.
8. Choi H, Lee H, Kim H. Fast detection and visualization of network attacks on parallel coordinates. *Comput Security*. 2009;28:276–288.
9. Choi H, Lee H. PCAV: Internet attack visualization on parallel coordinates. In: Proceedings of the Information and Communications Security, Vol. 3783. 2005:454–466.
10. Cheng KO, Law NF, Siu WC, Liew AWC. Biclusters visualization and detection using parallel coordinate plots. *Comput Model Life Sci*. (CMLS 07) 2007;952:114–123.
11. Lange B, Rey H, Vasques X, Puech W, Rodriguez N. Visualization assisted by parallel processing. In: Proceedings of the *Parallel Processing For Imaging Appl*, San Francisco, CA. 2011:7872.
12. Wang X, Medasani S, Marhoon F, Albazzaz H. Multidimensional visualization of principal component scores for process historical data analysis. *Ind Eng Chem Res*. 2004;43:7036–7048.
13. Chung KL, Zhuo N. Graph-based visual analytic tools for parallel coordinates. In: Proceedings of the Advances In Visual Computing, Pt Ii., Las Vegas, NV 2008:5359.
14. Zhou H, Yuan X, Qu H, Cui W, Chen B. Visual clustering in parallel coordinates. *Comput Graph Forum*. 2008;27:1047–1054.
15. Artero A, de Oliveira M, Levkowitz H. Uncovering clusters in crowded parallel coordinates visualizations. *IEEE Comp Soc Visual Graph TC*. 2004;81–88.
16. Hauser H, Ledermann F, Doleisch H. Angular brushing of extended parallel coordinates. Presented at the IEEE Symposium on Information Visualization (INFOVIS 2002), Boston, MA, Oct 28–29, 2002:127–130.
17. Albazzaz H, Wang X. Historical data analysis based on plots of independent and parallel coordinates and statistical control limits. *J Process Control*. 2006;16:103–114.
18. Valle S, Li W, Qin S. Selection of the number of principal components: the variance of the reconstruction error criterion with a comparison to other methods. *Ind Eng Chem Res*. 1999;38:4389–4401.
19. Macgregor J, Kourti T. Statistical process-control of multivariate processes. *Control Eng Pract*. 1995;3:403–414.
20. Wise B, Gallagher N. The process chemometrics approach to process monitoring and fault detection. *J Process Control*. 1996;6:329–348.
21. Dunia R, Qin S, Edgar T, McAvoy T. Identification of faulty sensors using principal component analysis. *AIChE J*. 1996;42:2797–2812.
22. Ge Z, Xie L, Song Z. A novel statistical-based monitoring approach for complex multivariate processes. *Ind Eng Chem Res*. 2009;48:4892–4898.
23. Liu X, Xie L, Kruger U, Littler T, Wang S. Statistical-based monitoring of multivariate non-Gaussian systems. *AICHE J*. 2008;54:2379–2391.
24. Lieftucht D, Yruger U, Irwin G. Improved reliability in diagnosing faults using multivariate statistics. *Comput Chem Eng*. 2006;30:901–912.
25. Zhou D, Li G, Qin SJ. Total projection to latent structures for process monitoring. *AIChE J*. 2010;56:168–178.
26. Urness T, Marrinan T, Johnson AR, Vitha MF. Multivariate visualization of chromatographic systems. Presented at the Conference on Visualization and Data Analysis, San Francisco, CA, Jan 24–25, 2011:7868.
27. Solka J, Wegman E, Marchette D. Data mining strategies for the detection of chemical warfare agents. *Stat Data Mining Knowledge Discov*. 2004:79–92.
28. Inselberg A. Visualization and data mining of high-dimensional data. *Chemom Intell Lab Syst*. 2002;60:147–159.
29. Qin S. Statistical process monitoring: basics and beyond. *J Chemom*. 2003;17:480–502.
30. Kourti T, Nomikos P, Macgregor J. Analysis, monitoring and fault-diagnosis of batch processes using multiblock and multiway PLS. *J Process Control*. 1995;5:277–284.
31. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol*. 1933;24:417.
32. Dunia R, Qin S. Subspace approach to multidimensional fault identification and reconstruction. *AIChE J*. 1998;44:1813–1831.
33. Dzyacky G, Carlson S. Distillation column flooding predictor—increase throughput, improve energy efficiency, and avoid flooding. Contributions to distillation troubleshooting and control. Presented at the AIChE Spring Conference, New Orleans, 2011.

Manuscript received Feb. 8, 2012, and revision received Apr. 13, 2012.