Monitoring of Mineral Processing Operations based on Multivariate Similarity Indices

L. Auret, C. Aldrich*

Department of Process Engineering, University of Stellenbosch, Private Bag XI, Matieland 7602, Stellenbosch, South Africa, Fax +27(21)8082059, *E-mail: ca1@sun.ac.za

Abstract: Multivariate process monitoring through covariance control charts considers changes in the relationships among process variables, but is limited by linearity assumptions. In this paper two nonlinear indicators of multivariate structure are considered, viz. mutual information and random forest proximity measures. Similarity matrices are constructed from data encapsulated by sliding windows of different sizes across the time series data associated with process operations. Diagnostic metrics reflect the differences between stationary base windows representative of normal operating conditions and test windows containing new process data. A case study in mineral processing shows that better results can be obtained with these nonlinear methods.

Keywords: Fault detection, machine learning, random forests, mineral processing

1. INTRODUCTION

Process monitoring has long since advanced from univariate methods such as Shewhart, cumulative sum and exponentially weighted moving average control charts for monitoring the means of a few key performance indicators (Shewhart, 1924; Page, 1954; Roberts, 1959). Extensions of these univariate charts to the multivariate monitoring of performance indicators and process variables allow broader consideration of process states and variable correlations (Hotelling, 1947; Crosier, 1988; Pignatiello & Runger, 1990; Lowry et al., 1992).

The fact that certain process disturbances may be characterized by a change in the multivariate structure of process variables has led to the development of multivariate control charts for monitoring covariance matrices. Yeh et al. (2006) provides a review of covariance monitoring methods. Sample covariance matrices can be monitored based on determinants (Alt & Bedewi, 1986); by conducting likelihood decomposition (Tang & Barnett, 1996), among others. Sample covariance matrices can be monitored based on determinants (Alt & Bedewi, 1986); by conducting likelihood decomposition (Tang & Barnett, 1996), among others.

A limitation of these covariance control charts stems from the linear nature of covariance as an indicator of intervariable relationships. When variables are not linearly correlated, covariance presents a limited indicator of relationship. To address this shortcoming, nonlinear measures of intervariable relationships, based on mutual information and unsupervised tree ensembles, are considered in this study.

Mutual information (Moddemeijer, 1989) can be seen as a more general measure of dependence between two variables than covariances or linear correlation coefficients.

Unsupervised learning with tree ensembles, such as random forests (Breiman, 2001), delivers a measure of similarity among training samples, based on the frequency of sample-pairs occupying similar local tree partitions (Breiman & Cutler, 2003). By calculating the forest ensemble similarities among process variables, a nonlinear approximation of intervariable relationships can be obtained.

Changes in these nonlinear similarity matrices over time may be quantified by the sum of squared deviations from a reference condition, and can be used as a monitoring metric for nonlinear process changes.

Two nonlinear multivariate similarity indices and a process monitoring method are developed in Section 2, with the application of these methods on an industrial coal flotation case study given in Section 3. Section 4 concludes with a general discussion and the conclusions of this study.

2. MULTIVARIATE SIMILARITY MONITORING

2.1 Mutual information

The mutual information of two random variables (say, x and y) quantifies the extent to which knowledge of the one variable captures information in the other, irrespective of the linearity of their relation. In terms of probabilities, mutual information represents the distance between the joint ($f_{xy}$) and marginal ($f_x$ and $f_y$) probability distributions of x and y (Moddemeijer, 1989):

$$I_{xy} = \sum_{x,y} f_{xy}(x,y) \log \frac{f_{xy}(x,y)}{f_x(x)f_y(y)}$$

(1)

With $I_{xy}$ the mutual information of x and y, $f_{xy}$ the probability density function of x and $f_{xy}$ the joint probability density function of x and y.

The mutual information for discrete data can be estimated by approximating probabilities through counting the co-occurrence of samples in a two-dimensional grid spanning x and y (Moddemeijer, 1989):
\[
\hat{I}_{xy} = \sum_{i,j} k_{ij} \log \frac{N_{k}}{k_{i}k_{j}}
\]

Here, \(k_{ij}\) is the number of samples in a specific cell \((i, j)\) in a specified two-dimensional grid; \(k_{i}\) is the sum of all samples in grid row \(i\); \(k_{j}\) is the sum of all samples in grid column \(j\); and \(N\) is the total number of samples.

Given a process data set \(X\) consisting of \(N\) samples and \(m\) variables, mutual information can be used to calculate the dependence between all possible variable-pair combinations to deliver an \(m \times m\) multivariate similarity matrix \(S\), with elements \(S_{ij}\) (\(i \neq j\)):

\[
S_{ij} = \hat{I}_{X_{i}X_{j}}
\]

With \(X_{i}\) indicating process variable \(i\).

### 2.2 Random forest proximities

Random forests are nonlinear regression or classification models consisting of ensembles of regression or classification trees, such that each tree depends on a random vector sampled independently from the data (Breiman, 2001). A classification tree (Breiman et al., 1993) divides the feature space into recursive binary partitions, allocating a fixed-value prediction to each partition. A random forest consists of \(K\) decision trees, where each tree is generated based on a random vector. With random forests, these random vectors describe bootstrapping of training samples and random selection of available node splitting variables for each tree and split. The detailed random forest construction algorithm is described by Breiman (Breiman, 2001).

Of interest in this study is the random forest proximity measure. Random forests can be used to construct proximity matrices for data on the basis of whether samples report to the same tree leaf nodes or not. A leaf node essentially spans a hyperrectangle in the input space; if two samples report to the same hyperrectangle, they must be proximal. The algorithm for constructing a proximity matrix (Breiman & Cutler, 2003) is summarized below:

Construct a random forest model on input data \(X\) with response \(y\). Create an empty similarity (proximity) matrix \(S\). For each tree \(k = 1\) to \(K\):

- For each sample combination \((i,j)\), determine the terminal nodes to which they report.
- If a sample combination \((i,j)\) report to the same terminal node, increase \(S_{ij}\) by one.
- Repeat for all possible sample combinations.

Scale the similarity matrix \(S\) by dividing by the number of trees \(K\); the similarity matrix is symmetric and positive definite, with entries ranging from 0 to 1, and diagonal elements equal to 1.

Where a data set has no response variable \(y\) (a so-called unsupervised task), random forest proximities between samples in \(X\) can still be determined by introducing a synthetic contrast class and casting the problem as a supervised learning task (Shi & Horvath, 2006). A similarity matrix \(S\) can then be calculated between the samples of the original unlabeled data, \(X\).

In this study, the random forest proximity approach is modified to deliver multivariate similarities by simply transposing the input data \(X\). The forest similarity matrix \(S\) now reflects the similarities between the variables of \(X\), and not the samples of \(X\).

The localized and disjoint nature of the decision tree algorithm (and by extension, the random forest algorithm) allows modeling of complex nonlinear data distributions. It is proposed that this nonlinear modeling capability would allow representation of nonlinear multivariate similarities that could indicate when the nature of such similarities changes.

### 2.3 Monitoring procedure

The proposed multivariate similarity methods can be applied in a moving window monitoring procedure. Given a multivariate process time series data set \(X\) with \(N\) samples and \(m\) variables, windows of size \(w\) are slid across the time series and a similarity diagnostic \(\delta\) calculated for each window.

Validation data matrices that represent normal operating conditions are evaluated to determine a similarity diagnostic threshold \(\delta_{w}\). The base and validation matrices are randomly drawn from the first 2\(w\) samples of \(X\). The assumption is therefore made that the first 2\(w\) samples of \(X\) represent normal operating conditions.

As sample size may affect both the mutual information and random forest similarity, the base, validation and test windows are selected to have the same sizes. Adjusting window sizes by correcting for sample size bias is an avenue for further research.

The monitoring algorithm is presented here:

- **Step 1: Determine a base similarity matrix \(S^{b}\)**
  From the first 2\(w\) samples of \(X\), randomly select \(w\) samples to create the base matrix \(X^{b}\). Calculate the multivariate similarity matrix \(S^{b}\) using one of the aforementioned methods.

- **Step 2: Determine a similarity diagnostic threshold \(\delta_{w}\)**
  From the first 2\(w\) samples of \(X\), randomly select \(w\) samples to create a validation matrix \(X^{v}\). Calculate the multivariate similarity matrix \(S^{v}\) using one of the aforementioned methods (covariance, mutual information or the random forest proximity). Calculate the similarity diagnostic \(\delta\) between \(S^{v}\) and \(S^{b}\) with (4):

\[
\delta = \sum_{i,w} (S_{ij}^{v} - S_{ij}^{b})^{2}
\]

Repeat this selection of a random validation matrix and similarity diagnostic calculation 100 times to obtain a distribution of validation similarity diagnostics. Assign \(\delta_{w}\) as the \((1-\alpha)\times100\) percentile of this distribution of validation similarity diagnostics. (The percentile approach avoids assumptions on the distribution of the similarity diagnostic).
In this work, \( \alpha = 0.01 \) is used, suggesting that one has 99% confidence that a detection is not a false alarm.

- **Step 3: Calculate scaled similarity diagnostics \( \delta^* \) for test windows**
  
  For iterations from \( i = 1 \) to \( N-w+1 \): Construct a test matrix \( X_T \) consisting of samples \( i \) to \( i+w-1 \) from \( X \). Calculate the test similarity matrix \( S_T \) for the current test matrix. Calculate the similarity diagnostic \( \delta \) between \( S_T \) and \( S_B \) with (5).

  \[
  \delta = \sum_{i \neq j} \left( S_{ij}^T - S_{ij}^B \right)
  \]

  (5)

  Calculate the scaled similarity diagnostic \( \delta^* \) from (6):

  \[
  \delta^* = \frac{\delta}{\delta_{\alpha}}
  \]

  (6)

- **Step 4: Investigate nature of abnormal event by inspecting similarity departures \( \epsilon_{ij} \)**

  Once the diagnostic series have been obtained, regions of interest can be inspected to identify which process variables showed the highest contribution to the similarity diagnostic. The contribution of a particular process variable pair (e.g. \( X_i \) and \( X_j \)) to the overall similarity diagnostic can be calculated with (7):

  \[
  \epsilon_{ij} = \frac{\left( S_{ij}^T - S_{ij}^B \right)}{\sum_{i \neq j} \left( S_{ij}^T - S_{ij}^B \right)}
  \]

  (7)

  To illustrate this multivariate similarity monitoring procedure, an industrial case study is now considered.

### 3. CASE STUDY: INDUSTRIAL COAL FLOTATION

#### 3.1 Process description

Real data from an industrial coal flotation process (Yang et al., 2007) is considered here, with a schematic of the process given in Fig. 1. The raw slurry stream is separated into a clean coal product stream and a tailings stream. Kerosene is used as a collector, while the foamer is a byproduct of an upstream process. Water is added to control the slurry concentration.

The key performance indicator for this separation process is the ash content of the clean coal product stream. Measuring ash content for both the raw slurry feed and the clean coal product is difficult and costly. An online coal slurry analyzer for periodically measuring the raw, clean and waste streams can produce ash content measures at time intervals ranging from 7.5 to 12.5 minutes, with clean coal ash content precision of 0.15% to 0.22% mean square deviation (Yang et al., 2001).

Ash content is related to the process variables, with a change in the process variable structure implying a possible change in the product stream ash content. This motivates the monitoring of the process variable structure as an indirect indication of the clean coal ash content and departure from a desirable process state.

---

**Fig. 1.** Schematic of the coal flotation process.

**Fig. 2.** Process variables and performance variable for coal flotation process. The change from desirable to undesirable conditions is indicated in the time series plot for ash content, at time step 218. A moving average (window size of 20 samples) for the ash content is shown. Five process variables are monitored, namely the flow of the raw slurry (\( X_1 \)), the density of the raw slurry (\( X_2 \)), the flow of water (\( X_3 \)), the flow of collector (\( X_4 \)) and the flow of foamer (\( X_5 \)). These variables, as well as the ash content of the clean coal product stream (the performance variable) are depicted in Fig. 2. From the performance variable, the process data can be split into two conditions, i.e. an initial desirable condition of
low ash content before sample index 218; and the later undesirable condition of generally high ash after sample index 218.

As mentioned in the previous section, the monitoring methodologies require the assumption that the first $2w$ samples of the process variables represent normal operating conditions. Window sizes of 25, 50 and 75 samples were used in this study, implying normal operating conditions represented by 50, 100 and 150 samples, respectively. These required sample sizes fall within the initial condition of low ash content.

### 3.2 Monitoring results

The results for the three variants of the multivariate similarity monitoring approaches (based on covariance, mutual information and random forest proximities) are presented in Figs. 3, 4 and 5; as well as in Table 1. Shown diagnostic sequences are the average diagnostics for 10 iterations.

From Figs. 3, 4 and 5, it is apparent that small window sizes give the least accurate monitoring results (in terms of showing high diagnostic metrics for the fault condition after 218 time steps). This may be due to the high variability of multivariate similarity indices based on small sample sizes. It is expected that the power of some statistically derived metric would decrease for small sample sizes. However, the statistical advantage of large sample sizes must be weighed against induced delays in detection and possible smoothing over rapid changes.

Fig. 3. Diagnostic sequences for multivariate similarity monitoring with covariance, for three window sizes $w$. A change occurred at 218 time steps.

Fig. 4. Diagnostic sequences for multivariate similarity monitoring with mutual information, for three window sizes $w$. A change occurred at 218 time steps.

Fig. 5. Diagnostic sequences for multivariate similarity monitoring with random forest proximities, for three window sizes $w$. A change occurred at 218 time steps. (1000 trees were constructed for each forest).

Fig. 5 suggests that an increase in the random forest proximity diagnostic threshold could markedly improve the overall detection accuracy of the random forest approach. An increase in the threshold would allow less incorrect detections in the normal operating conditions region, while not affecting the correct detections in the fault region.

An interesting observation for the random forest diagnostic sequences for larger window sizes (Fig. 5) is that a marked increase in the diagnostic starts after approximately 180 time steps. The commencement of the abnormal state was defined at 218 time steps, based on expert knowledge. Three factors may explain the early onset diagnostic increase: Firstly, the diagnostic is incorrect, and has no relation to the presence or otherwise of an abnormal state. The high values of the diagnostic for the duration of the indicated abnormal state empirically negate this. Secondly, the definition of the start of the abnormal state may be inaccurate. And thirdly, due to process lags, the input variables (on which the diagnostic is based) may indicate a change in the process before the performance variable (unseen by the diagnostic calculation) exhibits a change.
The success of abnormal event detection can be expressed in terms of false and missing alarm rates (FAR and MAR, respectively), given post priori information on abnormal event occurrence. False alarms occur when detections are made where no known abnormal event is present, while missing alarms occur where no detections are made where a known abnormal event occurred. For this case study, the post priori abnormal event classification is made in terms of classifying samples before time index 218 as indicating normal operating conditions, and classifying samples after time index 218 as constituting a fault condition. On the basis of this classification, false and missing alarm rates are given in Table 1. It is noted that the interpretation of false and missing alarm rates is dependent on the accuracy of the post priori characterization of normal / fault conditions.

<table>
<thead>
<tr>
<th></th>
<th>(w = 25)</th>
<th>(w = 50)</th>
<th>(w = 75)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAR (Covariance)</td>
<td>0.844</td>
<td>0.808</td>
<td>0.887</td>
</tr>
<tr>
<td>MAR (Covariance)</td>
<td>0.133</td>
<td>0.105</td>
<td>0.121</td>
</tr>
<tr>
<td>FAR (Mutual)</td>
<td>0.599</td>
<td>0.719</td>
<td>0.592</td>
</tr>
<tr>
<td>MAR (Mutual)</td>
<td>0.544</td>
<td>0.270</td>
<td>0.432</td>
</tr>
<tr>
<td>FAR (Forest)</td>
<td>0.880</td>
<td>0.862</td>
<td>0.669</td>
</tr>
<tr>
<td>MAR (Forest)</td>
<td>0.230</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

From Table 1, the mutual information approach delivers the lowest false alarm rates. However, the mutual information approach also has the highest missing alarm rates.

The random forest approach generally showed the lowest missing alarm rates (Table 1), albeit with the highest false alarm rates. However the false alarm rates for the random forest approach could be decreased by adjusting its diagnostic threshold (based on further understanding of the threshold generation, as discussed previously).

In the current approach, the diagnostic threshold is based on the 99% percentile of a distribution of validation diagnostics, based on randomly sampled validation matrices. The effect of adjusting the diagnostic threshold on the random forest proximity FAR and MAR is shown in Fig. 6.

From Fig. 6, as the diagnostic threshold increases, FAR decreases and MAR increases. This shows the trade-off of Type I and Type II errors. The equivalence points (where FAR is equal to MAR, for the same window sizes) provide a performance comparison measure for the three techniques, as given in Table 2. From Fig. 6, the random forest proximity equivalence points show improvement (lower FAR and MAR) as the window size increases, suggesting that larger sample sizes for the proximity matrices allow better detection.

Table 2. False alarm rates and missing alarm rates at equivalence points for multivariate similarity diagnostics applied to the coal flotation case study (based on varying the diagnostic threshold).

<table>
<thead>
<tr>
<th></th>
<th>(w = 25)</th>
<th>(w = 50)</th>
<th>(w = 75)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAR / MAR (Covariance)</td>
<td>0.651</td>
<td>0.575</td>
<td>0.704</td>
</tr>
<tr>
<td>FAR / MAR (Mutual)</td>
<td>0.552</td>
<td>0.479</td>
<td>0.451</td>
</tr>
<tr>
<td>FAR / MAR (Forest)</td>
<td>0.839</td>
<td>0.210</td>
<td>0.085</td>
</tr>
</tbody>
</table>

From Table 2, the random forest proximity diagnostic performs best, excluding the smallest window size, where all three techniques perform poorly.

To aid the interpretation of the nature of detected abnormal events, the pairwise similarity departures can be inspected, as calculated with (7). Fig. 7 shows this for the random forest proximity measure (window size of 75), comparing a base matrix spanning time steps 1 to 75 with a test matrix spanning time steps 350 to 424.

From Fig. 7, it is apparent that a change in the relation between process variables 3 and 4 accounts for the largest fraction of the total change in the multivariate similarity structure. Process variable 3 represents the flow rate of water to the flotation process, while process variable 4 indicates the flow of collector to the flotation process. These variables are intricately related to the flotation process through density, viscosity, froth stability, mass pull and other process characteristics. Expert knowledge may be employed once...
variables have been identified to interpret an abnormal event and aid in process recovery.

4. DISCUSSION AND CONCLUSIONS

In this work, two nonlinear approaches have been presented for the monitoring of multivariate similarity structure of process data, and these two approaches were compared to a linear approach on a coal flotation case study.

The random forest proximity approach showed the most promise in terms of successfully detecting a post priori indicated fault condition. The false alarm rate of the random forest approach, although high, may be reduced by modification of the diagnostic threshold without sacrificing correct detections for the fault condition.

The mutual information approach did not show good results for the coal flotation case study. However, it is postulated that the no-free lunch theorem applies to these nonlinear monitoring methods as well, viz. certain nonlinear metrics are successful for certain data distributions, with no overall superior technique for all possible data distributions. The existence of a diversity of nonlinear data distributions motivates the exploration of a diversity of nonlinear metrics.

The multivariate similarity structure approach can further benefit process interpretation, in that the contributions of paired variable similarities to the overall diagnostic can be inspected.

REFERENCES


