Supporting Information

Generalized Polynomial Chaos based Fault Detection and Classification for Nonlinear Dynamic Processes

Yuncheng Du[†], Thomas A. Duever[‡], Hector Budman[†*]

[†] Department of Chemical Engineering, University of Waterloo, Waterloo, ON, Canada N2L 3G1

[‡] Department of Chemical Engineering, Ryerson University, Toronto, ON, Canada, M5B 2K3

[*] Corresponding author: Tel.: +1 519 888 4567 x 36980; Fax: + 1 519 746 4979. Email: hbudman@uwaterloo.ca (H. Budman)

The objective of this case study is to compare the gPC model based fault detection and classification method with the empirical model based methods for process monitoring. The principal component analysis (PCA) is used for comparison [3].

One of the most standard methods consists of constructing a single PCA model and defines regions in the lower dimensional space which classify whether a particular fault has occurred. Let us assume the matrix X is used to store measurements for all operating modes (mean values), and then the sample covariance matrix S can be calculated as:

$$S = \frac{1}{n\text{-}1} X^T X = V \Lambda V^T \tag{S.1}$$

, where the diagonal matrix Λ contains the nonnegative real eigenvalues of decreasing magnitude. The matrix Λ can be used to optimally capture the variations of the data in X, and the loading vectors P corresponding to the first a largest singular values can be then calculated.

Using the sample covariance matrix S and the loading vectors P, the maximum score discriminant [24] can be used to estimate the likelihood that an observation $\boldsymbol{x}$ is the operating mode i, which can be calculated as:

$$f_i(x) = \frac{1}{2}(\boldsymbol{x} - \bar{\boldsymbol{x}}_i)^T P(P^T S_i P)^{-1} P^T (\boldsymbol{x} - \bar{\boldsymbol{x}}_i) + \ln(p_i) - \frac{1}{2}\ln[\det(P^T S_i P)] \tag{S.2}$$

$$\bar{\boldsymbol{x}}_i = \frac{1}{n_i} \sum_{\boldsymbol{x}_j \in \chi_i} \boldsymbol{x}_j \tag{S.3}$$

, where $\bar{\boldsymbol{x}}_i$ is the mean vector for operating mode i, n_i is the number of measurements in operating mode i, χ_i is the set of vectors x_j which belong to the operating mode i, and S_i is the sample covariance matrix for operating mode i.

The score discriminant can also be used for multiple PCA models [24]. Assuming the PCA models retain the important variations in discriminating between the faults (operating modes), and observations $\boldsymbol{x}$ is classified as being in the operating mode i with the maximum score discriminant:

$$f_i(x) = \frac{1}{2}\boldsymbol{x}^T P_i \sum\nolimits_{a,i}^{-2} P_i^T \boldsymbol{x} + \ln(p_i) - \frac{1}{2}\ln[\det(\sum\nolimits_{a,i}^{-2})] \tag{S.4}$$

, where P_i is the loading matrix for the operating mode i, $\sum_{a,i}$ is the diagonal matrix for the operating mode i, and p_i is the overall likelihood of the operating mode i.

For comparison, the fault detection and classification algorithms defined in eq S.3 and eq S.4 are compared with the *Level-1 algorithm* developed in Section 3.2 when the system is operating at steady states. For the model calibration with eq S.3 and eq S.4, 100 measurements for each operating mode are used, while ~81 measurements for each operating mode are used for the gPC model calibration with eq 11. The number of step changes of the unknown input (x_{A0}) among the 3 mean values in the *ML-PRS* is 300 for the model calibration with PCA. Thus a slightly larger number of measurements were selected for the calibration of the PCA algorithm as compared to our proposed gPC approach.

Three scenarios are considered: (i) measurements collected in the absence of measurement noise and variation on the feed mass fraction x_{A0}; (ii) measurements collected with measurement noise but no stochastics variation on x_{A0}; and (iii) both measurement noise and uncertainty on x_{A0} are considered. Table S.1 shows the result of Fault Classification Rate (*FCR*) for these three scenarios.

Table S.1 *FCR* with *PCA model* (steady state measurements)

x_{A0}	S.3			S.4		
	Case i	*Case ii*	*Case iii*	*Case i*	*Case ii*	*Case iii*
0.65	0.99	0.98	0.83	0.99	0.99	0.88
0.75	1	0.85	0.72	1	0.88	0.76
0.85	1	0.93	0.85	0.99	0.90	0.84
Average	0.997	0.92	0.80	0.993	0.923	0.827

In Table S.1, the variation on x_{A0} follows the same assumption as done for the gPC model and 1% measurement noise is used for simulations. To comply with the assumption that the system is operated around a fixed mean value with perturbations, the classification efficiency is investigated using the measured quantities before a switch between means occurred (see inset Figure 1 (b)-A). The measurements denote that the system is operating at steady state with constant mean values. It can be seen that the variation on x_{A0} and the measurement noise show strong influence on the classification of faults. As compared to the results in Table 7, the *FCR* is ~10 percent points lower than the gPC model based *Level-1 algorithm*. An explanation for the difference is that the principal component analysis (PCA) is a linear dimensionality reduction method. When the data components have nonlinear dependencies, PCA may require a larger dimensional representation than would be found by a nonlinear technique. Additionally, comparing Case-ii to Case-iii, the classification rate decreased by ~10 percent points, when the uncertainty on feed mass fraction x_{A0} is considered. One may argue that extra data is required for the model calibration with the PCA method to increase accuracy. The use of more training measurements may improve the classification rate but would increase the computational burden. The proposed gPC based method both addresses the nonlinearity by explicitly using a nonlinear model and necessitates less data, since it directly predict PDF profiles of the variables used for detection.