Journal of Process Control 71 (2018) 90–102

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Sparse canonical variate analysis approach for process monitoring

Qiugang Lu[a,c], Benben Jiang[b,c,*], R. Bhushan Gopaluni[a], Philip D. Loewen[d], Richard D. Braatz[c,*]

[a] Dept. of Chemical and Biological Engineering, The University of British Columbia, Vancouver, BC, V6T 1Z3, Canada
[b] Dept. of Automation, Beijing University of Chemical Technology, Beijing 100029, China
[c] Dept. of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, MA, 02139, USA
[d] Dept. of Mathematics, The University of British Columbia, Vancouver, BC, V6T 1Z3, Canada

ARTICLE INFO

Article history:
Received 26 July 2017
Received in revised form 14 August 2018
Accepted 19 September 2018
Available online 11 October 2018

Keywords:
Process monitoring
Fault detection and identification
Canonical variate analysis
Contribution plot
Tennessee Eastman process

ABSTRACT

Canonical variate analysis (CVA) has shown its superior performance in statistical process monitoring due to its effectiveness in handling high-dimensional, serially, and cross-correlated dynamic data. A restrictive condition for CVA is that the covariance matrices of dependent and independent variables must be invertible, which may not hold when collinearity between process variables exists or the sample size is small relative to the number of variables. Moreover, CVA often yields dense canonical vectors that impede the interpretation of underlying relationships between the process variables. This article employs a sparse CVA (SCVA) technique to resolve these issues and applies the method to process monitoring. A detailed algorithm for implementing SCVA and its formulation in fault detection and identification are provided. SCVA is shown to facilitate the discovery of major structures (or relationships) among process variables, and assist in fault identification by aggregating the contributions from faulty variables and suppressing the contributions from normal variables. The effectiveness of the proposed approach is demonstrated on the Tennessee Eastman process.

1. Introduction

Process faults refer to abnormal operations of the process such as process parameter drifts, sensor malfunctions, and sticky valves. If undetected and uncompensated, faults can result in loss of equipment, process efficiency, product quality, or life, and/or can harm the environment [1]. Rapid detection of occurrence of process faults (known as fault detection) and associated identification of faulty variables (termed as fault identification) have become imperative tasks for industrial processes, especially for large-scale processes that are increasingly integrated and involve a large number of strongly correlated process variables [2,3]. Widespread implementation of information-based technologies in manufacturing industries has generated large quantities of process data which have boosted the development and application of data-based fault detection and identification techniques. Data-based process monitoring techniques have been developed from multivariate control charts and dimensionality reduction techniques to multivariate time-series model and state-space model approaches.

Classical multivariate control charts include multivariate versions of Shewhart charts [4], cumulative sum (CUSUM) charts [5], and exponentially weighted moving average (EWMA) charts [6]. The fully dimensional versions of these methods are only suitable when the process has a small number of variables with moderate extent of cross-correlation among variables. The presence of large-scale processes whose data may contain redundant information has motivated the utilization of dimensionality reduction techniques such as principal component analysis (PCA) [4] and partial least squares (PLS) [7]. A prerequisite for these methods to yielding satisfactory performance of process monitoring is the lack of temporal (aka serial) correlation of variables, which is rarely met in modern chemical processes that are featured by slow dynamics and are increasingly of high sampling frequency. To address such process systems, multivariate time-series modeling [8] and dynamic PCA/PLS [9] have been proposed to handle the serially and jointly correlated process data. In the former approach, multivariate time-series models such as vector autoregressive model (VAR) and vector autoregressive moving average (VARMA) models have been developed with process monitoring often based on residuals (assumed to be independent and identically distributed) retained from one-step-ahead prediction with these acquired models. Obtaining such

* Corresponding authors at: Richard D. Braatz at Massachusetts Institute of Technology, Cambridge, MA 02139, USA; Benben Jiang at Beijing University of Chemical Technology, Beijing 100029, China
E-mail addresses: jiangbb@mail.buct.edu.cn, bbjiang@mit.edu (B. Jiang), braatz@mit.edu (R.D. Braatz).

https://doi.org/10.1016/j.jprocont.2018.09.009
0959-1524/

multivariate models is an expensive task, especially when the dimension is high, and such models typically have identifiability problems [10]. For the latter approach, lagged values of process variables are stacked together with current values and conventional PCA/PLS is applied to the augmented variable vector. Such dynamic PCA/PLS methods are effective in recovering dynamic models when the noise level is low. For moderate or high noise levels, dynamic dimensionality reduction methods cannot guarantee to give an accurate and minimal dynamic model for the process [9]. To circumvent this problem, in recent decades, state-space models, particularly based on canonical variate analysis (CVA), have become the mainstream in time-series modeling for statistical process monitoring [11]. The CVA state-space realization technique estimates the states by maximizing the cross-correlation between past input-output data and a window of future outputs. An advantage of using CVA is its computational efficiency that admits a solution by solving a generalized singular value decomposition (SVD). CVA can be applied to actual large-scale processes involving many variables that are both strongly autocorrelated and cross-correlated.

CVA-based fault detection, identification, and diagnosis have attracted extensive attention from the research community [12–14]. A drawback of CVA is the lack of sparsity in the canonical vectors, which hinders the intuitive interpretation of canonical loadings. That is, canonical variates are linear combinations of all variables. Moreover, to implement CVA, the inverse of the sample covariances of past inputs and outputs as well as that of the future outputs must exist [1,15]. When the sample size is small relative to (or less than) the number of variables, or collinearity between a subset of variables occurs, these sample covariance matrices become highly ill-conditioned or even singular. Traditional CVA is no longer suitable for process monitoring in such scenarios. A remedy is to use the canonical ridge [16] to replace those covariance matrices, i.e., by adding penalty terms to the diagonal of those covariance matrices to make them invertible. Although this approach mitigates the invertibility issue, the resultant canonical vectors are still dense. In addition to the interpretation problem associated with dense canonical vectors, in the fault identification stage, dense canonical vectors sum contributions to faulty status from all variables, thus rendering the faulty variables less distinguishable from the rest. All these considerations motivate the development of a sparse CVA (SCVA) method that can address poor conditioning in sample covariances, facilitate better interpretations of canonical vectors, and promote the identification of faulty variables after a fault is detected.

Sparse models for dimensionality reduction have emerged in recent years. Zou et al. [17] proposed an algorithm for sparse PCA by formulating PCA as a regression and adding a Lasso penalty to achieve sparsity in the principal component loadings. Other sparse PCA algorithms have been reported in [18,19]. Chun and Keleş [20] studied the sparse PLS method that is suitable for the circumstance with a large number of variables and a small number of samples. In the context of sparse canonical correlation analysis (CCA), Parkhomenko et al. [21] considered a sparse SVD method to derive sparse canonical vectors. Witten et al. [22] proposed a penalized matrix decomposition approach that unifies sparse PCA and sparse CCA into an optimization with sparsity constraints on parameters. Waaijenborg et al. [23] and Wilms and Croux [15] extend the alternating least-squares approach for CCA to sparse CCA by incorporating the elastic net or Lasso penalties. Sparse dimensionality reduction methods have not been extensively investigated in the area of process monitoring. Gajjar et al. [24] and Gao et al. [25] consider the use of sparse PCA from [17] for fault detection and diagnosis, in which it is shown that determining the sparsity of principal component loadings involves a tradeoff between attaining sparsity and maximizing the explained variance. A sparse global-local preserving projections method is reported in [26] that can maintain both global and local structures in the data. Such a structure-preserving approach aids the discovery of meaningful correlation between variables and greatly improves the interpretability of transformation vectors. Although these papers show that sparse versions of dimensionality reduction methods have significantly improved interpretability than their dense counterparts, these reported sparse models inherit the disadvantages of their associated dense methods in dealing with dynamic data from large-scale continuous processes. That is, it is well established that CVA provides much more accurate input-output descriptions and much more effective fault detection performance for dynamic processes than PCA, PLS, and their dynamic extensions, especially for large-scale processes with highly autocorrelated and cross-correlated variables [12–14]. As such, a sparse version of the CVA-based fault detection method would be expected to inherit both the advantages of sparse methods over dense methods, and the advantages of CVA over alternative dimensionality reduction methods. These combined advantages motivate this work. With such motivation, this article proposes a SCVA method that aims at keeping the merits of CVA in handling high-dimensional, serially, and jointly correlated data, while having the advantages of sparse canonical vectors in interpreting the process and promoting the identification of faulty variables.

The rest of this article is organized as follows. Section 2 briefly revisits canonical variate analysis. The proposed sparse CVA monitoring approach is developed in Section 3, where a guideline on selecting proper sparsity parameters is described, and the statistics for fault detection as well as contribution charts for fault identification based on sparse CVA are provided. The effectiveness of the proposed approach is demonstrated in the Tennessee Eastman process in Section 4, followed by conclusions in Section 5.

2. Canonical variate analysis revisited

Given two sets of random variables, canonical variate analysis is a dimensionality reduction method that seeks the maximum correlation between linear combinations of each of these two sets of variables [27]. These linear combinations are known as the *canonical variates* and the corresponding correlations are denoted as *canonical correlations*. Considering process input and output vectors $\boldsymbol{x} \in R^m$ and $\boldsymbol{y} \in R^n$, covariance matrices $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{yy}$ and cross-covariance matrix $\boldsymbol{\Sigma}_{xy}$, the canonical vectors $\boldsymbol{J} \in R^{m\times m}$ and $\boldsymbol{L} \in R^{n\times n}$, that maximize canonical correlations satisfy the conditions [28]

$$\boldsymbol{J}\boldsymbol{\Sigma}_{xx}\boldsymbol{J}^{\mathrm{T}} = \boldsymbol{I}_{\bar{m}}, \quad \boldsymbol{L}\boldsymbol{\Sigma}_{yy}\boldsymbol{L}^{\mathrm{T}} = \boldsymbol{I}_{\bar{n}}, \quad \boldsymbol{J}\boldsymbol{\Sigma}_{xy}\boldsymbol{L}^{\mathrm{T}} = \boldsymbol{D} = \mathrm{diag}\,(\gamma_1, \ldots, \gamma_r, 0, \ldots, 0), \tag{1}$$

where $\gamma_1 \geq \ldots \geq \gamma_r$ are the canonical correlations, r is the rank of $\boldsymbol{\Sigma}_{xy}$, $\bar{m}$ and $\bar{n}$ are the rank of $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$ respectively, and $\boldsymbol{I}_{\bar{m}} \in R^{m\times m}$ denotes a diagonal matrix with the first $\bar{m}$ diagonal elements as one and the other diagonal elements as zero. Variables in the vector of canonical variates $\boldsymbol{c} = \boldsymbol{J}\boldsymbol{x}$ are mutually uncorrelated with a covariance matrix $\boldsymbol{\Sigma}_{cc} = \boldsymbol{I}_{\bar{m}}$. The same holds for the vector of canonical variates $\boldsymbol{d} = \boldsymbol{L}\boldsymbol{y}$. Moreover, variables in $\boldsymbol{c}$ and $\boldsymbol{d}$ are pairwise correlated. A standard algorithm to compute the projection matrices $\boldsymbol{J}$ and $\boldsymbol{L}$ involves a singular value decomposition (SVD) of the form

$$\boldsymbol{\Sigma}_{xx}^{-1/2}\boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1/2} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\mathrm{T}}, \tag{2}$$

where $\boldsymbol{J} = \boldsymbol{U}^{\mathrm{T}}\boldsymbol{\Sigma}_{xx}^{-1/2}$, $\boldsymbol{L} = \boldsymbol{V}^{\mathrm{T}}\boldsymbol{\Sigma}_{yy}^{-1/2}$, $\boldsymbol{D} = \boldsymbol{\Sigma}$, the unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ can be interpreted as rotation operations such that variables in $\boldsymbol{c}$ and $\boldsymbol{d}$ are only pairwise correlated, and $\boldsymbol{\Sigma}_{xx}^{-1/2}$ and $\boldsymbol{\Sigma}_{yy}^{-1/2}$ are scaling matrices to ensure that elements in $\boldsymbol{c}$ and $\boldsymbol{d}$ have unit variance. An implicit assumption is that the covariance matrices $\boldsymbol{\Sigma}_{xx}$

and $\boldsymbol{\Sigma}_{yy}$ are invertible, which is invalid when certain variables in $\boldsymbol{x}$ or $\boldsymbol{y}$ are collinear. Moreover, in practice, $\boldsymbol{\Sigma}_{xx}$ and $\boldsymbol{\Sigma}_{yy}$ are replaced by their respective sample covariance matrices. Numerical issues may arise if variables in $\boldsymbol{x}$ (or) are close to collinear, or the sample number N is small relative to the number of variables $m+n$. These issues invoke modifications of canonical variate analysis, such as penalized CVA [16] and SCVA as presented in the next subsection.

Canonical variate analysis can be viewed as an implementation of canonical correlation analysis [29] typical in multivariate statistics for time series modeling. Akaike first proposed to unitize CVA in the context of stochastic realization theory and system identification of ARMA models. CVA was further extended to the state-space modeling of time series data by Larimore [28]. The classical form of state-space model is given as

$$\boldsymbol{x}(t+1) = \boldsymbol{A}\boldsymbol{x}(t) + \boldsymbol{B}\boldsymbol{u}(t) + \boldsymbol{v}(t), \tag{3}$$

$$\boldsymbol{y}(t) = \boldsymbol{C}\boldsymbol{x}(t) + \boldsymbol{D}\boldsymbol{u}(t) + \boldsymbol{E}\boldsymbol{v}(t) + \boldsymbol{w}(t), \tag{4}$$

where $\boldsymbol{A},\boldsymbol{B},\boldsymbol{C},\boldsymbol{D},\boldsymbol{E}$ are system matrices of appropriate dimensions, $\boldsymbol{x}(t) \in R^d$ is a d -order state vector, $\boldsymbol{v}(t)$ and $\boldsymbol{w}(t)$ are sequences of white noise with zero mean and constant covariances, and $\boldsymbol{u}(t) \in R^{n_u}$ and $\boldsymbol{y}(t) \in R^{n_y}$ are input and output signals that are typically measured by sensors in industrial processes. n_u and n_y are the dimensions of input and output signals, respectively. Repeated iterations of (3) and (4) imply that the values of the state up to any point in time is linearly related to past inputs $\{\boldsymbol{u}(t-1), \boldsymbol{u}(t-2), \ldots\}$ and past outputs $\{\boldsymbol{y}(t-1), \boldsymbol{y}(t-2), \ldots\}$. In CVA, finite input and output sequences are used to estimate the state vector. Specifically, define

$$\boldsymbol{p}(t) = \left[\boldsymbol{y}^{\mathrm{T}}(t-1), \ldots, \boldsymbol{y}^{\mathrm{T}}(t-l), \boldsymbol{u}^{\mathrm{T}}(t-1), \ldots, \boldsymbol{u}^{\mathrm{T}}(t-l)\right]^{\mathrm{T}}, \tag{5}$$

$$\boldsymbol{f}(t) = \left[\boldsymbol{y}^{\mathrm{T}}(t), \ldots, \boldsymbol{y}^{\mathrm{T}}(t+h)\right]^{\mathrm{T}}. \tag{6}$$

Assuming the state order is d, the CVA seeks a projection matrix $\boldsymbol{J}_d$ to map the past information vector to the d -dimensional "memory" vector $\boldsymbol{m}(t) = \boldsymbol{J}_d\boldsymbol{p}(t)$ in order to minimize the prediction error

$$\mathbb{E}\{[\boldsymbol{f}(t) - \hat{\boldsymbol{f}}(t)]^{\mathrm{T}} \Lambda^{\dagger} [\boldsymbol{f}(t) - \hat{\boldsymbol{f}}(t)]\},$$

where $\hat{\boldsymbol{f}}(t)$ is the linear optimal forecast of $\boldsymbol{f}(t)$ based on the current memory, i.e., $\hat{\boldsymbol{f}}(t) = \boldsymbol{\Sigma}_{fm}\boldsymbol{\Sigma}_{mm}^{-1}\boldsymbol{m}(t)$, where $\boldsymbol{\Sigma}_{fm}$ is the covariance between $\boldsymbol{f}(t)$ and $\boldsymbol{m}(t)$, and $\boldsymbol{\Sigma}_{mm}$ is defined similarly. The positive semidefinite weighting matrix Λ reflects the relative importance among output variables. It is shown by [13] that such a problem can be addressed by canonical variate analysis. Substituting $\boldsymbol{p}(t)$ and $\boldsymbol{f}(t)$ to $\boldsymbol{x}$ and $\boldsymbol{y}$ in (1), respectively, the projection matrices $\boldsymbol{J}_d$ can be obtained by solving the SVD as in (2) with $\boldsymbol{J}_d = \boldsymbol{U}_d^{\mathrm{T}}\boldsymbol{\Sigma}_{pp}^{-1/2}$ and $\boldsymbol{U}_d$ consists of the first d columns of the unitary matrix $\boldsymbol{U}$. Note that the number d of canonical variates can be chosen to be greater than the state order in the minimal realization of the true system. The first d canonical variates (also known as *canonical states*) are acquired from

$$\boldsymbol{x}_d(t) = \boldsymbol{J}_d\boldsymbol{p}(t). \tag{7}$$

The canonical state vector $\boldsymbol{x}_d(t)$ is an estimate of the true state vector, upon which various performance monitoring statistics can be established.

Remark. The lags, h and state order d constitute the tuning parameters that are crucial in determining the performance of the canonical state estimation. A practical approach for selecting these tuning parameters applies ARX model structures with different orders to the process data and chooses the order that yields the minimal Akaike information criterion [28] as the candidate. The lags l and h can be determined according to the orders of the optimal ARX model, and d is determined based on the state order of its minimal realization. Increasing the lags l and h can enhance the accuracy of the models given by CVA, but too many lags will lead to overfitting and decrease the robustness of process monitoring techniques. In addition, more data would be required for the parameter estimation and the computational cost would increase accordingly. As pointed out in [14], experience shows that usually $h = 1$ *or* 2 suffices for process monitoring in practice.

3. The proposed sparse canonical variate analysis based approach for fault monitoring

3.1. Sparse canonical variate analysis (SCVA) method

As discussed above, when the sample covariance matrices of $\boldsymbol{\Sigma}_{xx}$, $\boldsymbol{\Sigma}_{yy}$ are singular or ill-conditioned, the conventional CVA method may deteriorate or fail due to the induced numerical issues. Moreover, CVA produces dense canonical vectors that combine all variables into a canonical variate. This fact impedes obtaining an intuitive interpretation about the structure of relations among underlying variables. SCVA arises in this context to ensure the feasibility of CVA and discover the major relationships between variables with sparse canonical vectors. To formulate SCVA, the first pair of canonical vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ from traditional CVA can be obtained by solving the optimization [22]:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{\beta} \quad \text{s.t. } \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\alpha} \le 1,\ \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{Y}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{\beta} \le 1, \tag{8}$$

where $\boldsymbol{X} \in R^{N\times m}$ and $\boldsymbol{Y} \in R^{N\times n}$ are standardized data matrices containing N samples of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. It is easy to verify that the optimal $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ always activate the inequality constraints and thus the unit variance constraint on canonical variates is satisfied. With SCVA, our objective is to maximize the correlation between linear combinations of $\boldsymbol{x}$ and $\boldsymbol{y}$ while restricting the canonical vectors to contain only a few nonzero elements. One approach is to add extra constraints to (8) that enforces the sparsity of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. A well-known option is the l_1 constraint that poses an upper bound on the sum of absolute values of entries in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. With this idea, (8) is re-formulated into its sparse form as

$$\begin{aligned}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}\boldsymbol{\beta} \quad &\text{s.t. } ||\boldsymbol{\alpha}||_2^2 \le 1,\ \left||\boldsymbol{\beta}|\right|_2^2 \le 1,\ ||\boldsymbol{\alpha}||_1 \le c_1,\\ &\left||\boldsymbol{\beta}|\right|_1 \le c_2,\end{aligned} \tag{9}$$

where c_1 and c_2 are two tuning parameters specifying the sparsity in $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The covariances of $\boldsymbol{X}$ and $\boldsymbol{Y}$ have been approximated by diagonal matrices, which have been shown to produce satisfactory results especially when the data are high-dimensional [30]. This approximation is assumed to hold throughout this article.[1] The optimization can be efficiently addressed via penalized matrix decomposition as proposed in [22]. Subsequent pairs of canonical vectors follow in the well-known deflation form in which $\boldsymbol{X}^T\boldsymbol{Y}$ is replaced by residuals from previous canonical vectors (refer to Algorithm 1 below). In the context of process input and output data, the above SCVA can be directly applied by substituting $\boldsymbol{x}$ and $\boldsymbol{y}$ with $\boldsymbol{p}(t)$ and $\boldsymbol{f}(t)$, respectively. The first pair of canonical vectors is acquired by solving

$$\begin{aligned}\max_{\boldsymbol{\alpha},\boldsymbol{\beta}} \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{P}^{\mathrm{T}}\boldsymbol{F}\boldsymbol{\beta} \quad &\text{s.t. } ||\boldsymbol{\alpha}||_2^2 \le 1,\ \left||\boldsymbol{\beta}|\right|_2^2 \le 1,\ ||\boldsymbol{\alpha}||_1 \le c_1,\\ &\left||\boldsymbol{\beta}|\right|_1 \le c_2,\end{aligned} \tag{10}$$

where $\boldsymbol{P} \in R^{(N-h-l)\times(n_y l + n_u l)}$ stacks past information $\boldsymbol{p}(t)$, $t = l+1, \ldots, N-h$, into a matrix and $\boldsymbol{F} \in R^{(N-h-l)\times(n_y h)}$ contains the future information. After the first pair of canonical vector arrives,

[1] This assumption is widely used in the literature to simplify the analysis, and extensive studies have shown the validity of this assumption; e.g., see [22,23].

the second pair of canonical vectors is derived simply by applying (10) to the residuals of $\boldsymbol{P}^{\mathrm{T}}\boldsymbol{F}$. Algorithm 1 is modified from [22] to be suitable for computing the projection matrices for $\boldsymbol{p}(t)$ and $\boldsymbol{f}(t)$. Before demonstrating the main algorithm, first define the soft-thresholding function to be $\boldsymbol{S}(\boldsymbol{a},\boldsymbol{c}) = sign(\boldsymbol{a})(|\boldsymbol{a}| - \boldsymbol{c})_+$, where $\boldsymbol{a}$ and $\boldsymbol{c}$ can be either vectors or scalars, $sign(\boldsymbol{a})$ and $|\boldsymbol{a}|$ respectively take the sign and absolute value of $\boldsymbol{a}$, and x_+ equals x if $x > 0$ and 0 otherwise. The main algorithm is shown below (for more details, refer to [22]).

3.2. Selection of sparsity penalty parameters c_1 and c_2

The selection of sparsity penalty parameters c_1 and c_2 plays a fundamental role in trading off between enforcing the sparsity of CVA vectors $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$, and maximizing the correlations $corr(\boldsymbol{X\alpha}, \boldsymbol{Y\beta})$. Such types of tradeoff have been extensively reported in the literature for a variety of sparse models, such as sparse PCA [17,24], sparse CCA [15], and sparse PLS [20]. Classical selection methods include the BIC and AIC criteria and cross-validation. This work employs the cross-validation strategy where the sparse canonical vectors are obtained from the training data and examined via the validation data. The averaged (or accumulated) cross-correlations in the validation data are used as the selection criterion and can be computed by applying trained pairs of canonical vectors to the validation set. The space of c_1 and c_2 are gridded according to their respective intervals and those values are chosen to yield maximum averaged correlations in the validation data. The value of c_1 is bounded below by 1 and above by $\sqrt{n_y l + n_u l}$ (see Appendix A for the proof). Similarly, the value of c_2 has a lower bound of 1 and upper bound $\sqrt{n_y h}$. For simplicity, this article chooses a unique sparsity tuning parameter c such that $c_1 = c_2 = c\sqrt{n_y l + n_u l}$ for both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$.

Algorithm 1: SCVA with penalized matrix decomposition

$\boldsymbol{Z}^1 \leftarrow \boldsymbol{P}^{\mathrm{T}}\boldsymbol{F}$

Outer Loop: For $k \in 1, \ldots, d$, where d is the number of canonical vectors

Initialize $\boldsymbol{v}$ to have unit l_2 norm. Repeat the following until convergence:

Inner Loop:
- $\boldsymbol{u} \leftarrow \boldsymbol{S}(\boldsymbol{Z}^k\boldsymbol{v}, \Delta_1)$ where $\Delta_1 = 0$ if it results in $\|\boldsymbol{u}\|_1 \le c_1$; otherwise, Δ_1 is chosen by a binary search such that $\|\boldsymbol{u}\|_1 = c_1$
- $\boldsymbol{v} \leftarrow \boldsymbol{S}((\boldsymbol{Z}^k)^{\mathrm{T}}\boldsymbol{u}, \Delta_2)$ where $\Delta_2 = 0$ if it results in $\|\boldsymbol{v}\|_1 \le c_2$; otherwise, Δ_2 is chosen by a binary search such that $\|\boldsymbol{v}\|_1 = c_2$

End Inner Loop

$\gamma_k \leftarrow \boldsymbol{u}^{\mathrm{T}}\boldsymbol{Z}\boldsymbol{v}$, $\boldsymbol{\alpha}^k \leftarrow \boldsymbol{u}$, $\boldsymbol{\beta}^k \leftarrow \boldsymbol{v}$. Update the residual $\boldsymbol{Z}^{k+1} \leftarrow \boldsymbol{Z}^k - \gamma_k \boldsymbol{u}\boldsymbol{v}^{\mathrm{T}}$.

End Outer Loop

Output: $\boldsymbol{J}_d \leftarrow [\boldsymbol{\alpha}^1, \ldots, \boldsymbol{\alpha}^d]^{\mathrm{T}}$, $\boldsymbol{L}_d \leftarrow [\boldsymbol{\beta}^1, \ldots, \boldsymbol{\beta}^d]^{\mathrm{T}}$, $\boldsymbol{D} \leftarrow \mathrm{diag}(\gamma_1, \ldots, \gamma_d)$

Remark. Similar to CVA, the proposed SCVA method also seeks the maximization of the correlations between past and future information vectors, except that a set of constraints are formulated to achieve the sparsity. Compared with other sparse dimensionality reduction methods, the SCVA reserves the proficiency of CVA in addressing the dynamic process data with large input-output dimensions and strong serial correlations. SCVA has slightly more complex computations due to its iterative algorithm presented above. However, in most process monitoring applications, the training stage is performed offline, in which case the extra computational cost of SCVA does not pose a concern.

3.3. SCVA-based statistics for fault detection

Two types of statistics are commonly used for process monitoring [14]. Hotelling's T^2 statistic measures the variations in the sparse canonical state space and is defined as

$$T_d^2 = \boldsymbol{x}_d^{\mathrm{T}}(t)\boldsymbol{\Lambda}^{-1}\boldsymbol{x}_d(t) \tag{11}$$

where $\boldsymbol{x}_d(t) = \boldsymbol{J}_d\boldsymbol{p}(t)$, $\boldsymbol{\Lambda}$ is the covariance matrix of canonical variates from training data, and $\boldsymbol{J}_d$ is the matrix of sparse canonical vectors obtained from Algorithm 1. The subscript d of T_d^2 is used in (11) to explicitly show that this statistic measures variations inside the state space. For conventional CVA, $\boldsymbol{\Lambda}$ is an identity matrix since the attained canonical variates are mutually uncorrelated. For SCVA, $\boldsymbol{\Lambda}$ may not be equal to an identity matrix due to the l_1 penalty in (10); i.e., the obtained canonical variates from SCVA are usually correlated. Given a level of significance α, the corresponding control limit of Hotelling's T^2 statistic is $T_{d,\alpha}^2 = \frac{d(N^2-1)}{N(N-d)}F_\alpha(d, N-d)$, where $F_\alpha(d, N-d)$ is the upper α percentile of the F distribution with degree of freedom d and $N-d$ [14]. The Q statistic measures the variations in the residual space. For non-orthogonal projection loadings, a common approach to obtain the optimal reconstructed $\hat{\boldsymbol{p}}(t)$ is through the pseudoinverse of the projection matrix, i.e., $\hat{\boldsymbol{p}}(t) = \boldsymbol{J}_d^{+}\boldsymbol{x}_d(t)$ [31] and the residual signal is then calculated as

$$\boldsymbol{r}(t) = (\boldsymbol{I} - \boldsymbol{J}_d^{+}\boldsymbol{J}_d)\boldsymbol{p}(t) \tag{12}$$

where $\boldsymbol{J}_d^{+} = (\boldsymbol{J}_d^{\mathrm{T}}\boldsymbol{J}_d)^{-1}\boldsymbol{J}_d^{\mathrm{T}}$ is the Moore-Penrose pseudoinverse of $\boldsymbol{J}_d$. The Q statistic is defined as

$$Q = \boldsymbol{r}^{\mathrm{T}}(t)\boldsymbol{r}(t). \tag{13}$$

A typical threshold for the Q statistic is given in Eq. 4.22 in [14] but such a threshold builds upon an assumption that the noise distribution is normal. In this work, a threshold for the Q statistic is determined based on the training data set. Specifically, given a level of significance α, the threshold Q_α is set in such way that a $(1-\alpha)$ portion of training samples are below the threshold. To evaluate the overall performance for a provided data example, the overall statistic $S_{overall}$ is defined to be the logical 'or' operation between T_d^2 and Q:

$$S_{\mathrm{overall}} = \begin{cases} 1, & \text{if } T_d^2 > T_{d,\alpha}^2 \text{ or } Q > Q_\alpha \\ 0, & \text{otherwise} \end{cases} \tag{14}$$

A test example is considered as faulty if either T_d^2 or Q violates their respective thresholds, i.e., if $S_{overall}$ returns one. In general, the T_d^2 statistic measures the status of states and a violation of the T_d^2 threshold indicates that the states are out of control. Exceeding Q_α normally implies changes in the characteristics of noise, or that new states have been created. The value of $S_{overall}$ assesses the overall health of process loops and is used to compare the performance of different monitoring techniques in this article.

Remark. An alternative approach to compute residuals for CVA is to use the last few rows of $\boldsymbol{J}$ in (2), which is known to be overly sensitive to the inversion of small values in $\boldsymbol{\Sigma}_{xx}^{-1/2}$ and leads to increased incidence of false alarms. Computing residuals in such a manner would also substantially increase the computational cost since it would require computing all loadings in $\boldsymbol{J}$ with Algorithm 1 instead of only the first d loadings.

3.4. SCVA-based contributions for fault identification

Once a fault is discovered, contribution plots are employed to identify the individual contribution from each variable to this faulty status. Although the canonical state $\boldsymbol{x}_d(t)$ itself cannot directly indicate the contribution from each variable, such information can be acquired from the projection matrix $\boldsymbol{J}_d$. For the state space in a CVA model at time t, the k th element $p_k(t)$ of new data $\boldsymbol{p}(t)$ has a contribution $c^d_{p_k}(t)$ computed as [1]

$$c^d_{p_k}(t) = \boldsymbol{x}_d^{\mathrm{T}}(t)\boldsymbol{\Lambda}^{-1}p_k(t)\boldsymbol{J}_{d,k} \tag{15}$$

where $\boldsymbol{J}_{d,k}$ is the k th column of the matrix $\boldsymbol{J}_d$. The contribution from each controlled variable and manipulated variable is more valuable, which involves, for a specific process variable, adding up all its past contributions as the signature of this variable's contribution. For example, the contribution of controlled variable $y_m(t)$, $m = 1, \ldots, n_y$, is expressed as

$$c^d_{y_m}(t) = \sum_{j=1}^{l} \boldsymbol{x}_d^{\mathrm{T}}(t)\boldsymbol{\Lambda}^{-1}y_m(t-j)\boldsymbol{J}_{d,m_j} \tag{16}$$

where m_j is the index of column of $\boldsymbol{J}_d$ that corresponds to variable $y_m(t-j)$. The contribution for each manipulated variable is computed in an analogous way.

In terms of the contribution for the residual space of a SCVA model, first define $\boldsymbol{J}_e = \boldsymbol{I} - \boldsymbol{J}_d^{+}\boldsymbol{J}_d$. The matrix $\boldsymbol{J}_e$ is likely to be sparse, under the condition that the number of variables is large while the canonical state order is relatively small and sparse. The expression of contributions for each variable (e.g., $y_m(t)$) in the residual space can be similarly deduced as

$$c^r_{\boldsymbol{y}_m}(t) = \sum_{l}^{j=1} \boldsymbol{r}^{\mathrm{T}}(t)y_m(t-j)\boldsymbol{J}_{e,m_j}. \tag{17}$$

As commented in [1], a higher contribution of a process variable indicates a more severe abnormal status of the underlying variable. A significant contribution of a variable based on state space usually signifies a larger deviation of relevant states with respect to those states in the normal operation stage. Faulty variables identified through residual space generally occur with the creation of new states in the system due to changes in the process or noise, and the original CVA model is no longer valid. Due to the possible numerical inaccuracies, a joint contribution plot based on state space and residual space contributions can reduce the incorrect identification of faulty variables. These three types of contribution plots are demonstrated in the next section.

4. Application to the Tennessee Eastman process

The Tennessee Eastman process (TEP) is a well-known benchmark process that is widely used to compare various fault detection and identification techniques. The TEP simulator was designed to provide sufficient simulation data that reflect the operation of the actual process with high fidelity. A diagram of the TEP is shown in Fig. 1. This process consists of five major operation units: a two-phase reactor, a condenser, a compressor, a vapor/liquid separator, and a stripper. A detailed description of the process model employed in the simulator as well as the plant-wide control structure is referred to [14] and the references therein. Table 1 lists the process variables. The process has 41 process output measurements, x_1 through x_{41}, consisting of 22 continuous process measurements, x_1 through x_{22}, and 19 composition measurements,

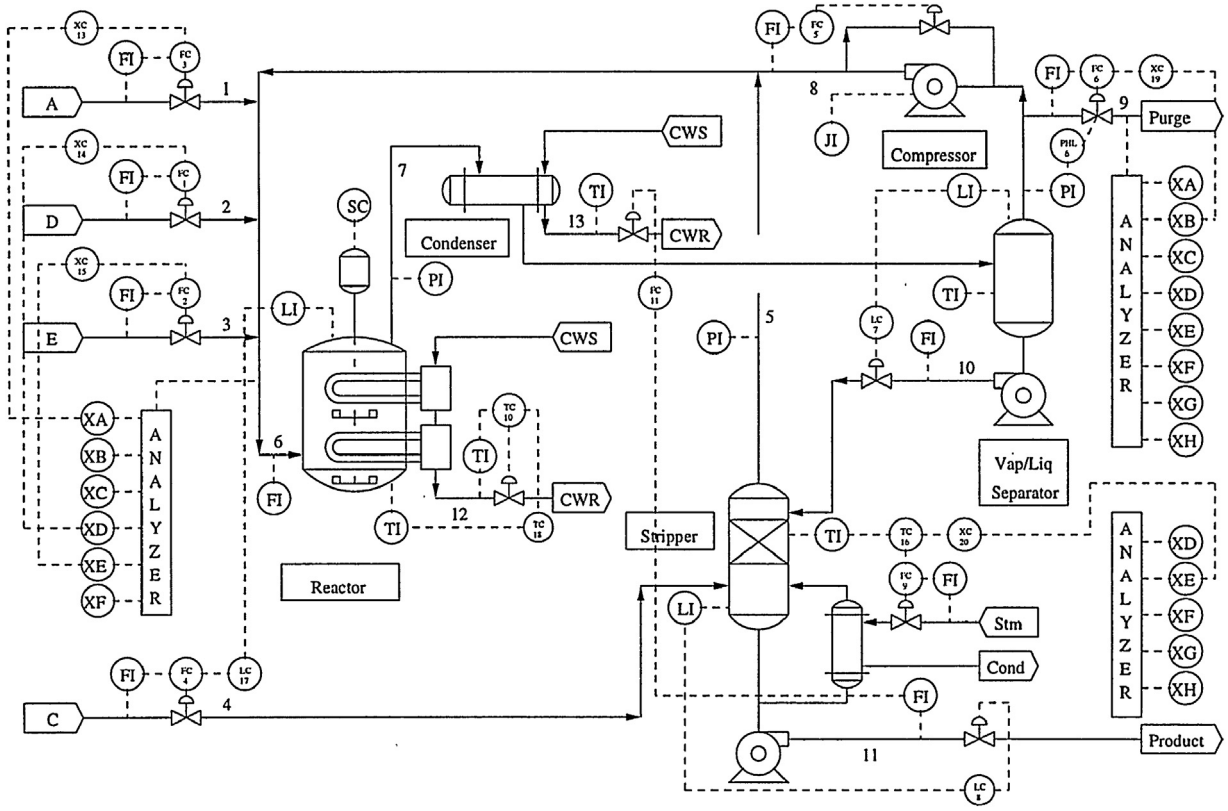

Fig. 1. Flow chart of the Tennessee Eastman process.

Table 1
Monitored variables in the Tennessee Eastman process [20].

ID	Variable description	ID	Variable description
x_1	A feed (Stream 1)	x_{27}	Component E (Stream 6)
x_2	D feed (Stream 2)	x_{28}	Component F (Stream 6)
x_3	E feed (Stream 3)	x_{29}	Component A (Stream 9)
x_4	A and C feed (Stream 4)	x_{30}	Component B (Stream 9)
x_5	Recycle flow (Stream 8)	x_{31}	Component C (Stream 9)
x_6	Reactor feed rate (Stream 6)	x_{32}	Component D (Stream 9)
x_7	Reactor pressure	x_{33}	Component E (Stream 9)
x_8	Reactor level	x_{34}	Component F (Stream 9)
x_9	Reactor temperature	x_{35}	Component G (Stream 9)
x_{10}	Purge rate (Stream 9)	x_{36}	Component H (Stream 9)
x_{11}	Product separator temperature	x_{37}	Component D (Stream 11)
x_{12}	Product separator level	x_{38}	Component E (Stream 11)
x_{13}	Product separator pressure	x_{39}	Component F (Stream 11)
x_{14}	Product separator underflow (Stream 10)	x_{40}	Component G (Stream 11)
x_{15}	Stripper level	x_{41}	Component H (Stream 11)
x_{16}	Stripper pressure	x_{42}	MV to D feed flow (Stream 2)
x_{17}	Stripper underflow (Stream 11)	x_{43}	MV to E feed flow (Stream 3)
x_{18}	Stripper temperature	x_{44}	MV to A feed flow (Stream 1)
x_{19}	Stripper stream flow	x_{45}	MV to total feed flow (Stream 4)
x_{20}	Compressor work	x_{46}	Compressor recycle valve
x_{21}	Reactor cooling water outlet temperature	x_{47}	Purge value (Stream 9)
x_{22}	Separator cooling water outlet temperature	x_{48}	Separator pot liquid flow (Stream 10)
x_{23}	Component A (Stream 6)	x_{49}	Stripper liquid product flow (Stream 11)
x_{24}	Component B (Stream 6)	x_{50}	Stripper steam valve
x_{25}	Component C (Stream 6)	x_{51}	Reactor cooling water flow
x_{26}	Component D (Stream 6)	x_{52}	Condenser cooling water flow

x_{23} through x_{41}. There are 12 manipulated (or input) variables, x_{42} through x_{52}. These process variables (52 variables in total) are used to validate the performance monitoring technique.

The training dataset contains 500 observations and is fault free. A sampling interval of 3 min is used to record these data. For the testing data, pre-programmed 21 faults are simulated to generate 21 faulty datasets corresponding to different faults encountered in practice. Moreover, an additional fault 0 (with no fault) testing data is available as the validation dataset. Each testing dataset has 960 samples, starting with no fault and then a fault is introduced after 160 samples (8 h). The process variables have a variety of units and scales. A normalization step is necessary for training, validation, and testing datasets before implementing any performance monitoring technique. After normalization, the observations of each variable have zero mean and unit variance.

4.1. Determining the sparsity parameter values

As discussed in Section 3.2, a common sparsity parameter c was selected for both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. A lower bound of c is $max\{1/\sqrt{n_y l + n_u l}, 1/\sqrt{n_y h}\}$ and an upper bound is one. This case study chooses lags $l = h = 2$, which are from an earlier study [14] that contains a detailed explanation on their optimal choice. The interval [0.12 0.8] on c was gridded with a step size of 0.02. SCVA Algorithm 1 was applied for each c to obtain a set of canonical vectors $\boldsymbol{J}_d$ and $\boldsymbol{L}_d$. For the state order of $d = 23$, the value of $c = 0.18$ gave the best averaged cross-correlation of d pairs of canonical variates on the validation data (see Fig. 2). With this selected sparsity parameter, the set of canonical vectors $\boldsymbol{J}_d$ and $\boldsymbol{L}_d$ was stored and implemented for fault detection and identification. This case study compares SCVA to the traditional CVA. For the selected parameters, the condition numbers of the sample covariance matrices of both $\boldsymbol{p}(t)$ and $\boldsymbol{f}(t)$ were very high. Poor conditioning implies that either the sample size is low relative to the number of variables or some of these variables are nearly collinear. Although the covariance matrices can be inverted within the accuracy (10^{-16}) of Matlab, such poor conditioning discourages the usage of traditional CVA. An observation that supports this statement is that some of the elements in the vectors $\boldsymbol{J}_d$ and $\boldsymbol{L}_d$ obtained from CVA were extremely large.

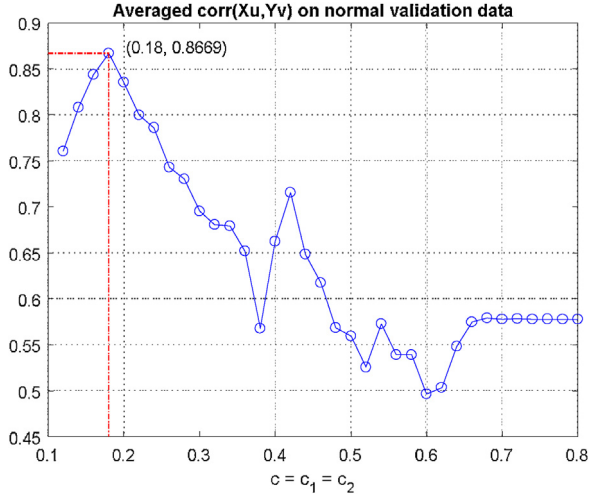

Fig. 2. Averaged cross-correlation under the validation data for different values of c.

The resultant inaccuracies in $\boldsymbol{J}_d$ and $\boldsymbol{L}_d$, due to the inaccuracies in the estimates of $\boldsymbol{\Sigma}_{pp}^{-1/2}$ and $\boldsymbol{\Sigma}_{ff}^{-1/2}$, leads to poorer performance of traditional CVA compared to SCVA, as discussed in the next section.

4.2. Fault detection

This section compares the fault detection performance of SCVA with that of conventional CVA. A key metric in evaluating fault detection performance is the missed detection rate. The missed detection rate is defined as the ratio of undetected faulty samples, by the T_d^2, Q, or $S_{overall}$ criteria, relative to the total number of faulty samples under a specific fault. With the selected sparsity parameter c, the specific missed detection rates for these three statistics under SCVA and CVA are shown in Table 2. For the overall statistic $S_{overall}$, SCVA had 40.4% (100(27.8–19.8)/19.8) lower missed fault detection rate than CVA. The missed fault detection rate of SCVA is at the same level or substantially lower than CVA for all of the faults except for Faults 5 and 21.

Table 2
Missed fault detection rates for 21 faults under the condition that ($l = 2,\ c_1 = 0.20,\ c_2 = 0.20,\ \lambda = 0.01$).

Fault	SCVA			CVA			CVA with ridge penalty		
	T_d^2	Q	$S_{overall}$	T_d^2	Q	$S_{overall}$	T_d^2	Q	$S_{overall}$
1	0.001	0.007	0.001	0	0.058	0	0.001	0.004	0.001
2	0.010	0.014	0.010	0.009	0.033	0.009	0.010	0.014	0.010
3	0.865	0.967	0.837	0.856	0.980	0.850	0.878	0.788	0.729
4	0.910	0.004	**0.004**	0.655	0.918	**0.634**	0.723	0.028	0.028
5	0.659	0.724	0.625	0	0	0	0.683	0.606	0.571
6	0.001	0	0	0	0	0	0	0	0
7	0.457	0	0	0.379	0.750	0.378	0.285	0	0
8	0.013	0.024	0.013	0.016	0.190	0.016	0.018	0.009	0.009
9	0.908	0.967	0.883	0.883	0.988	0.882	0.901	0.809	0.758
10	0.082	0.626	0.078	0.173	0.802	0.172	0.289	0.289	0.222
11	0.812	0.247	0.218	0.555	0.836	0.527	0.614	0.231	0.222
12	0	0.008	0	0	0.033	0	0.006	0.001	0.001
13	0.043	0.063	0.043	0.040	0.113	0.038	0.043	0.040	0.040
14	0.852	0	0	0	0.018	0	0	0	0
15	0.774	0.932	0.745	0.755	0.982	0.752	0.809	0.747	0.693
16	0.034	0.824	0.031	0.152	0.641	0.152	0.375	0.360	0.262
17	0.302	0.069	0.060	0.074	0.166	0.055	0.115	0.067	0.058
18	0.080	0.102	0.079	0.093	0.112	0.093	0.095	0.077	0.072
19	0.028	0.888	0.024	0.709	0.940	0.686	0.887	0.791	0.730
20	0.124	0.567	0.115	0.220	0.865	0.220	0.398	0.299	0.267
21	0.393	0.683	0.393	0.369	0.723	0.366	0.404	0.535	0.399
Average:	0.350	0.367	**0.198**	0.283	0.483	**0.278**	0.359	0.271	**0.242**
Std:	0.370	0.397	0.303	0.319	0.413	0.315	0.342	0.311	0.285

The bold values in row 4 show the case of Fault 4 where SCVA presents a superior performance than traditional CVA. The bold values in the averaged missed detection rate verify that SCVA shows better fault detection performance on average over all 21 faults than the other two methods.

SCVA relying only on the T_d^2 statistic would yield a higher missed detection rate compared with CVA, which can be explained by the fact that the canonical state space of a CVA model mainly captures the significant predictive relationships among variables. Pursuit of sparsity in such models is generally at the price of sacrificing the prediction accuracy and, as a result, SCVA is less sensitive than CVA in detecting faults in the state space. In other words, only relatively large changes in the state space can be detected by SCVA. Fortunately, extensive published results (e.g., [14]) have shown that most faults are better detected in the residual space by the Q statistic, in which SCVA showed 31.6% (100(0.483–0.367)/0.367) better fault detection performance compared to CVA. The residual space stores variations not captured by the state-space model, such as noise or other weak relationships between variables. If a fault brings minor changes to the process or only affects the noise characteristics, the fault can be easily detected by the Q statistic of SCVA.

In summary, SCVA loses some fault detection sensitivity in the state space which is compensated by increased sensitivity of its Q statistic. The high missed detection rate of CVA is largely due to the large condition numbers of $\boldsymbol{\Sigma}_{pp}$, $\boldsymbol{\Sigma}_{ff}$. To verify this sensitivity, CVA was repeated with a small ridge penalty term $\lambda = 0.01$ added to these covariance matrices (as shown by the last column of Table 2). The use of such regularized covariance matrices reduced the overall missed detection rate to 24.2%, which is an improvement although not as good as SCVA (cf. Table 2). This case study thus shows the suitability of applying SCVA when $\boldsymbol{\Sigma}_{pp}$ and $\boldsymbol{\Sigma}_{ff}$ are nearly singular.

Now consider the detection of Fault 4, which is known to be very challenging to detect [12] and for which SCVA was especially effective. Fault 4 introduces a step change at the 160th sample to the reactor cooling water inlet temperature which causes a step change directly in the manipulated variable x_{51}, which is the reactor cooling water flow (Fig. 3a). Consequently, a sudden increase in the

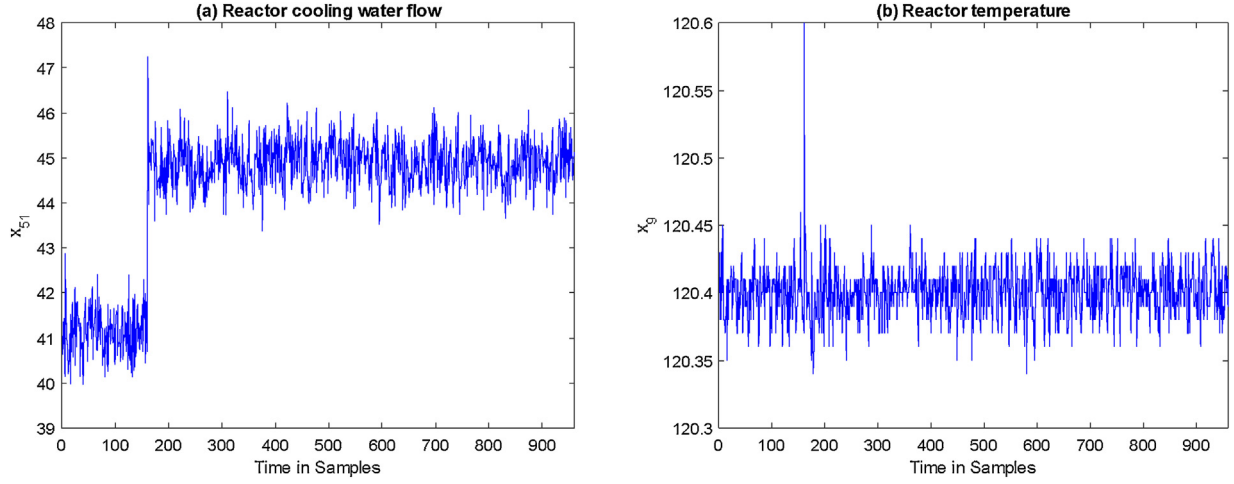

Fig. 3. Effects of Fault 4 on (a) reactor cooling water flow x_{51} and (b) reactor temperature x_9.

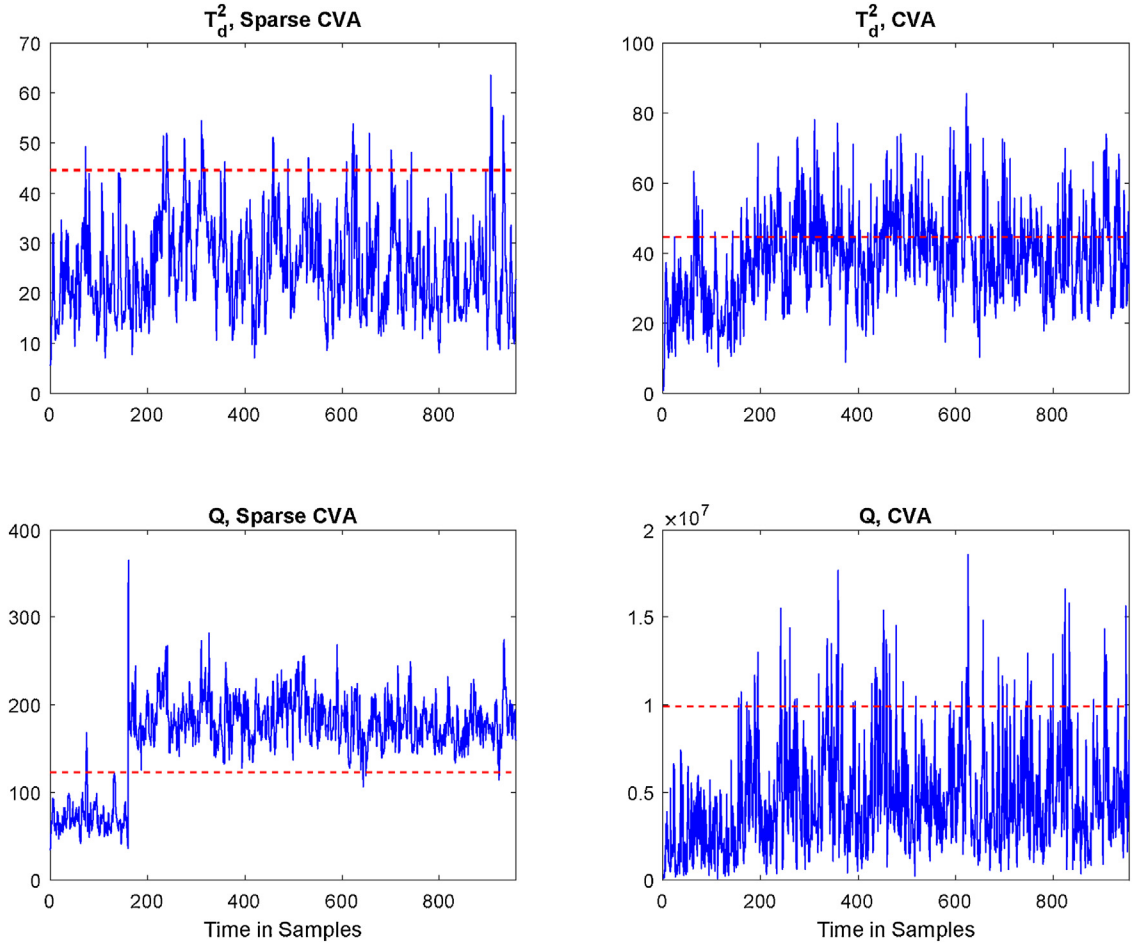

Fig. 4. Fault detection results for Fault 4 with SCVA (left) and traditional CVA (right). The thresholds are shown as horizontal dashed lines.

reactor temperature (x_9) appears after the 160th sample but then is quickly compensated by a control loop (Fig. 3b). Fault 4 is known to be challenging to detect both in the state space by the T_d^2 statistic and in the residual space by the Q statistic, as seen by Table 4 in [12] which could only detect Fault 4 by using a T_r^2 statistic in the residual space. The residual spaces for T_r^2 and Q are constructed in a different manner with the residual space for T_r^2 requiring the last few canonical vectors of $\boldsymbol{J}$, which can be sensitive to perturbations in the testing data for some faults [12]. From Table 4 in [12], the missed fault detection rates for CVA under both the T_d^2 and Q statistics are high for Fault 4, which agrees with the results in Table 2. In contrast, although the missed fault detection rate for the SCVA-based T_d^2 statistic is high, its Q statistic provides a persistent indication of faulty status with nearly zero missed fault detection rate (see bottom left plot of Fig. 4). The Q statistic of SCVA showed a much higher sensitivity than that of CVA, which is consistent with the above analyses.

To further compare the performance of SCVA, CVA, and ridge CVA, the false positive rates (FPR) from normal testing data for the three methods are reported in Table 3. According to Table 3, SCVA demonstrates similar performance as CVA, which both have less FPR than ridge CVA. Moreover, the detection delays in samples from these methods for each fault are shown in Table 4. For the T_d^2 statistic, for most faults, SCVA has similar performance as the other methods, with slightly worse performance for Faults 4, 15, and 17. This observation is consistent with previous observation that the T_d^2 statistic for SCVA is less sensitive to the faults. The Q statistic for SCVA yields similar performance as ridge CVA and much better performance than CVA except for Faults 9 and 15. Collectively, in terms of FPR and detection delay, SCVA presents comparable or superior performance than the other two methods.

4.3. Interpretation of canonical vectors

CVA is unsuitable for scenarios with a limited number of samples or the presence of collinear variables. Although CVA with ridge penalty can resolve this issue, a drawback of this approach is that dense canonical vectors are still obtained that combine all variables, which is not beneficial for interpretation of the canonical vectors. SCVA can not only directly handle poorly conditioned covariance matrices, but also produces sparse loadings such that the discovery of process knowledge becomes straightforward. This is the main

Table 3
False positive rates (FPR) for testing data under the condition that ($l = 2,\ c_1 = 0.20,\ c_2 = 0.20,\ \lambda = 0.01$).

Method	SCVA			CVA			CVA with ridge penalty		
Statistics	T_d^2	Q	$S_{overall}$	T_d^2	Q	$S_{overall}$	T_d^2	Q	$S_{overall}$
Average FPR	0.051	0.018	0.065	0.061	0.012	0.066	0.052	0.079	0.110

Table 4
Detection delays (samples) for 21 faults under the condition that ($l = 2,\ c_1 = 0.20,\ c_2 = 0.20,\ \lambda = 0.01$).

Fault	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
SCVA T_d^2	2	9	7	58	1	1	1	10	1	19	1	1	35	1	91	1	34	17	1	62	9
CVA T_d^2	1	6	6	12	1	1	1	12	1	18	6	1	24	1	66	10	19	12	10	78	20
Ridge T_d^2	2	10	6	12	1	1	1	12	1	18	6	2	26	1	88	10	19	14	12	82	20
SCVA Q	7	12	15	1	1	1	1	14	59	15	6	2	41	1	241	1	23	61	10	89	285
CVA Q	1	22	25	12	1	1	1	17	8	5	10	2	10	2	118	17	19	60	14	107	292
Ridge Q	4	12	14	1	1	1	1	17	2	13	6	2	7	1	92	14	24	55	10	84	257

advantage of implementing SCVA for fault detection and identification compared with CVA. More importantly, sparse loadings can strengthen the contributions of major variables relevant to the faults, so that the resultant fault identification becomes more accurate compared with using dense canonical vectors. The latter will be discussed in the next subsection. This subsection focuses on unveiling the structure (or relationships between variables) of a process, particularly the TEP, by using SCVA.

For illustrative purposes, a case study is considered in which process variables are selected to contain only the first 22 measurement variables (x_1 to x_{22}) and the 11 manipulated variables (x_{42} to x_{52}). The composition measurements are excluded to simplify the analysis. The past and future lags are set to one, with $\boldsymbol{p}(t)$ and $\boldsymbol{f}(t)$ stacking related variables in the same fashion as (5)–(6). As a rule of thumb, a sparser CVA model tends to preserve more fundamental variables in $\boldsymbol{p}(t)$ that can predict $\boldsymbol{f}(t)$ with larger prediction errors. For sparsity parameter $c = 0.28$ and state order $d = 16$, the structure of the set of canonical vectors is shown in Fig. 5.

The horizontal axis shows the process variables with the first 22 variables being measurements and the rest being manipulated variables. The vertical axis displays each canonical vector of SCVA. Each loading vector is sparse and dominated by a small number of nonzero elements. Similar to the physical interpretation provided by sparse loading in residual space as in [32], the large nonzero elements in SCVA loadings typically represent physical or control links between corresponding process variables. The reason is that the dominant process variables with large loading usually represent that they are highly correlated due to their connections, either through physical or control links. The major connections between variables discovered from Fig. 5 are summarized in Table 5.

Table 5
Physical [·] and control (·) links between variables.

Loading #	Nonzero elements	Loading #	Nonzero elements
1	$[x_7, x_{13}, x_{16}]$	9	$[x_{20}, x_{27}]$
2	(x_{31}, x_{19}, x_{18})	10	$[x_{27}, x_{20}, x_{11}]$
3	(x_1, x_{25})	11	(x_{18}, x_{31})
4	(x_{23}, x_2)	12	(x_{16}, x_{28}, x_{31})
5	$[x_{16}, x_7, x_{13}]$	13	$[x_{22}, x_7]$
6	(x_{28}, x_{10})	14	$[x_{16}, x_{31}]$
7	(x_{32}, x_9)	15	$[x_7, x_{13}]$
8	(x_{18}, x_{19})	16	(x_{18}, x_{31})

Most of the relationships uncovered by SCVA in Table 2 represent actual connections (either physical or control) between these variables, except loading 12 that singles out x_{16}, x_{28}, x_{31} as major

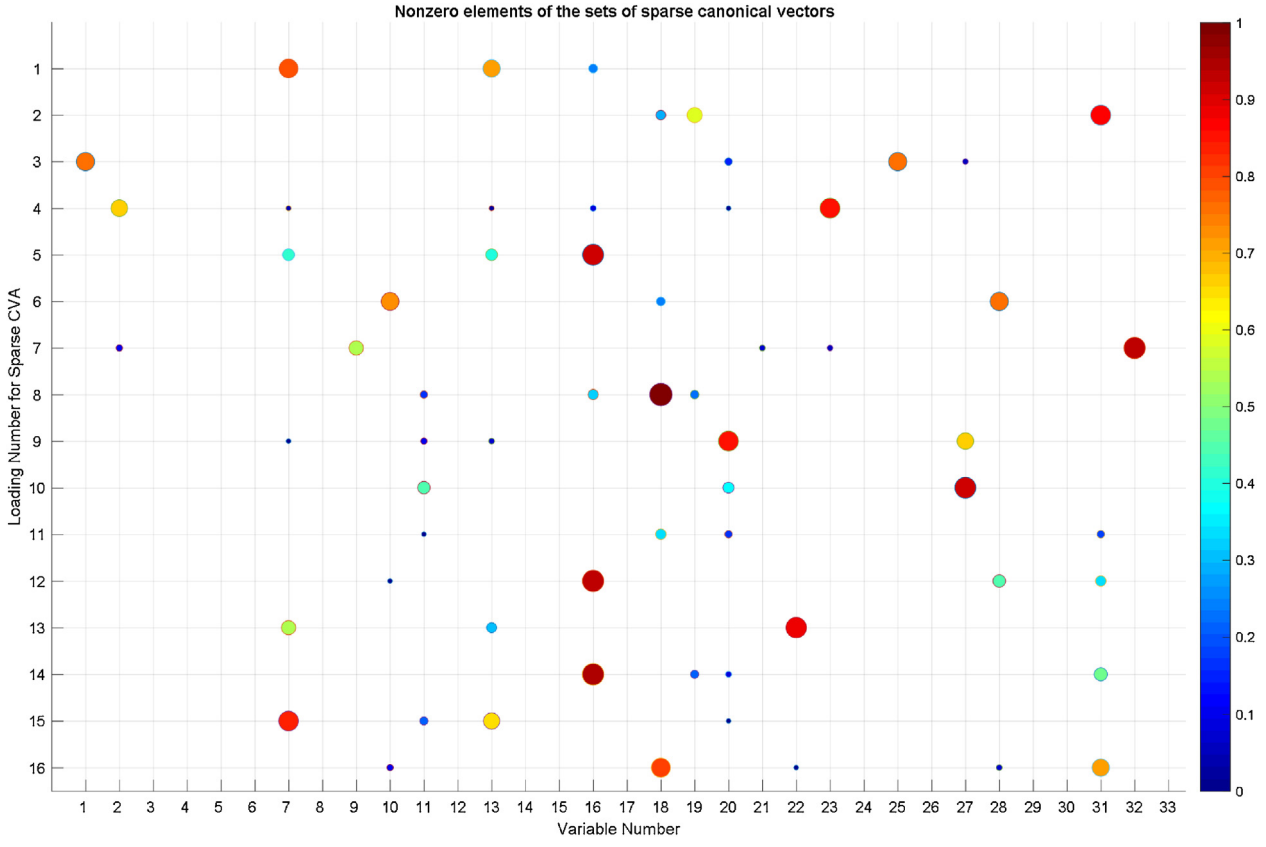

Fig. 5. Sparsity structure of the set of canonical vectors. The size and color of each point represent the absolute value of a nonzero element in a canonical vector.

variables. A summary of the uncovered knowledge from these sparse loadings is shown below:

- $[x_7, x_{13}, x_{16}]$: the pressure measurements for reactor, separator and stripper due to their physical connections.
- (x_{31}, x_{19}, x_{18}) : the causal-effect relation between stripper steam flow x_{19} and stripper temperature x_{18}, as well as the control connection between stripper steam flow x_{19} and stripper steam valve x_{31}.
- (x_1, x_{25}): control link between A feed x_1 and A feed flow x_{25}.
- (x_{23}, x_2): control link between D feed x_2 and D feed flow x_{23}.
- (x_{28}, x_{10}): control link between purge rate x_{10} and purge valve x_{28}.
- (x_{32}, x_9): control link between reactor temperature x_9 and reactor cooling water flow x_{32}.
- (x_{18}, x_{19}): control link between the stripper temperature x_{18} and stripper steam flow x_{19}.
- $[x_{20}, x_{27}]$: physical link between compressor work x_{20} and compressor recycle valve x_{27}.
- $[x_{22}, x_7]$: cause-effect relation between condenser cooling water outlet temperature x_{22} and reactor pressure x_7.
- $[x_{16}, x_{31}]$: cause-effect relationship between stripper steam valve x_{31} and stripper pressure x_{16}.

It is observed that some variables appear multiple times in different loadings. Recall that the loadings produced by SCVA are not orthogonal and the canonical variates are not uncorrelated. As a result, variables that are left out are not significant in minimizing the prediction error. Process knowledge discovery through sparse models have been reported in [25,26], mainly by sparse PCA. Their work obtains similar results as here, but with some minor differences that are caused by two reasons. First, the objective in their work [25,26] is to obtain sparse principal components to explain as much variance in the data as possible, whereas this article is concentrated on attaining pairs of sparse vectors to achieve maximum canonical correlations (or minimal prediction error) between two sets of data. Second, to acquire main relationships between variables [25,26], specified initial conditions carefully based on prior knowledge of the process. In contrast, here the initial sparse canonical vectors are chosen randomly, as initial conditions are typically not arbitrarily specifiable in a large-scale industrial process.

These discovered relationships can be used to gain better insights into the process by observing the most crucial variables for fault detection and identification. This information is useful in fault identification through contribution plots. SCVA can reinforce the contributions of faulty variables and weaken the contributions from other variables, as compared with dense CVA. This point is demonstrated in the next subsection.

4.4. Contribution plots based on SCVA

Contribution plots are a popular technique to identify faulty variables that are most relevant to causes of a fault. Large contributions from certain variables under a faulty scenario indicate that

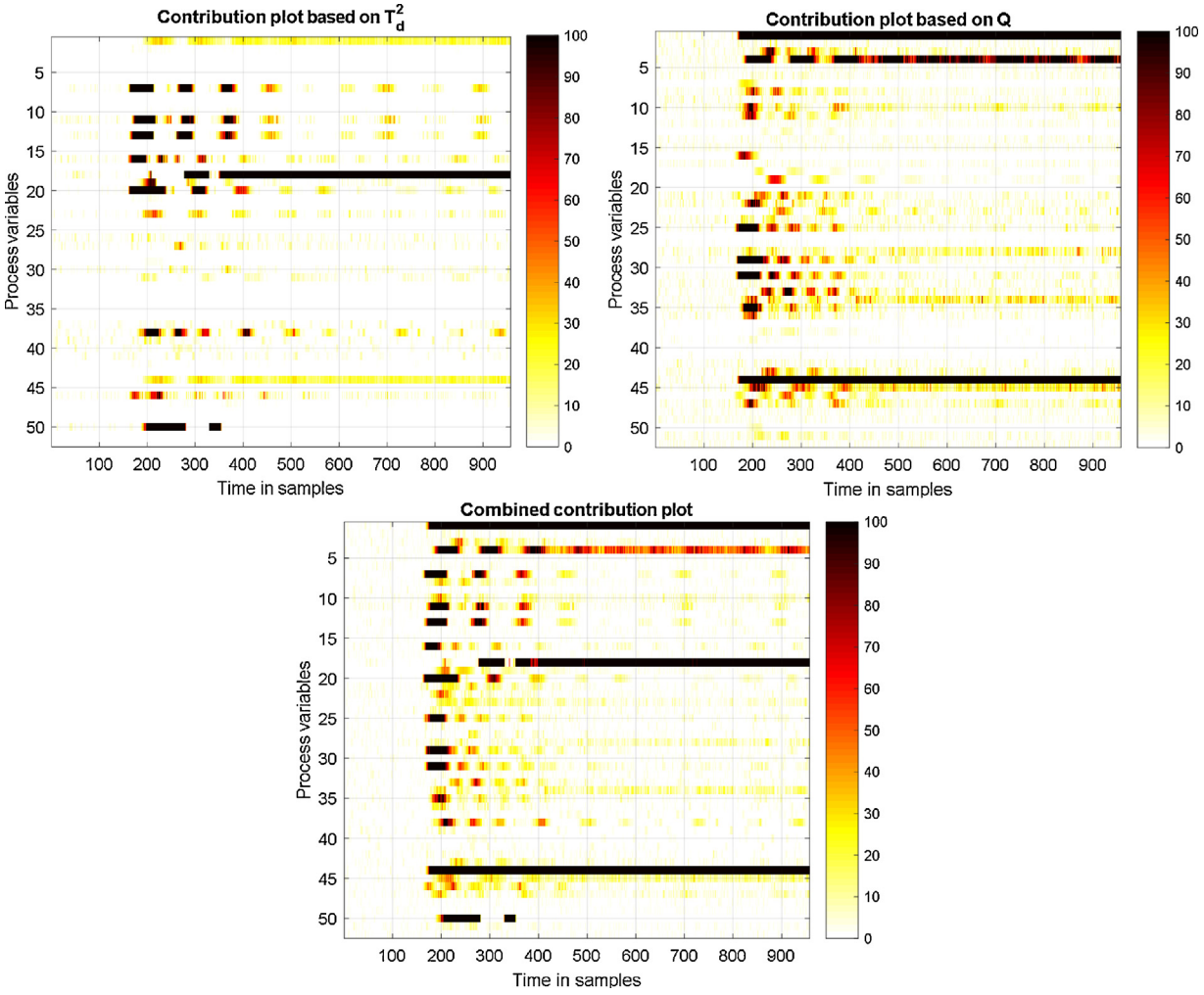

Fig. 6. Contribution plots based on T_d^2, Q, and combined SCVA statistics (the fault occurs at the 160th sample).

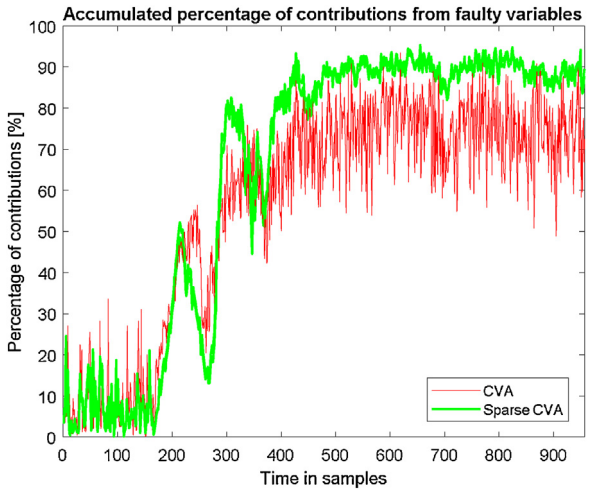

Fig. 7. Accumulated contributions from the faulty variables for Fault 1 identified by SCVA.

those variables are most likely to be the causes of the fault. An intuitive demonstration of faulty variables varying with time is the 2D color plot of contributions of all variables [33]. In such plots, the horizontal axis shows the time and the vertical axis represents all process variables, whereas the color in each grid indicates the value of contribution. This section demonstrates that SCVA can be better than CVA in singling out faulty variables.

Fault 1 is a step-type fault that causes a change after the 160th sample in the A / C feed ratio in Stream 4. This fault brings an increase in the C composition and a decrease in the A composition. As a result, the A composition decreases in Stream 5, which causes an increase in the A composition in Stream 1 due to the corrective action of control loops. A subsequent impact is variations in flowrate and compositions in Stream 6, which changes the reactor level and in turn perturbs the flowrate in Stream 4 which results from the control connections between level sensor (x_8) and feed flow valve (x_{45}). Variations in the C composition in Stream 4 due to Fault 1 also cause changes in the E composition due to material balance in reactions. As a result, Fault 1 affects the compositions of A, C, E and eventually propagates to many other variables and products. Thus, Fault 1 is expected to be relatively easy to detect, which agrees with the low missed detection rate shown in Table 2.

SCVA-based contribution plots relying on T_d^2, Q, and combined statistics are shown in Fig. 6. The combined contribution plot in Fig. 6c for variable $y_m(t)$ is computed by averaging the contributions from T_d^2 and Q : $c_{y_m}(t) = \left(c_{y_m}^d(t) + c_{y_m}^r(t)\right)/2$. Many variables show large contributions right after Fault 1 is introduced. As the control loop makes efforts to compensate for Fault 1, the contributions of most variables settle to steady-state values (see Fig. 6c) by the 400th sample. The variables that tend to give large contributions even after the 400th sample are x_1, x_4, x_{18}, and x_{44}. These identified faulty variables are similar as reported in [34]. Most faulty variables (except x_{18}) are identified through Fig. 6b, which is due to the reason explained in Section 4.2 that the Q statistic tends to have higher sensitivity. The faulty variables should not be identified based solely on the contribution plot from the Q statistic since that statistic includes noise that can be averaged out in the combined contribution plot.

The faulty variables x_1, x_4, x_{18}, and x_{44} correspond to the A feed in Stream 1, total feed in Stream 4, stripper temperature, and A feed flow valve in Stream 1. Since the stripper has a direct connection with Stream 4, it is reasonable that some of its properties are heavily associated with Fault 1. Moreover, the compensation from control loops drastically impacts Stream 1, causing x_1 and x_{44} to be the most evident reflections of Fault 1.

Fig. 6 verifies the effectiveness of using SCVA to extract faulty variables. In order to show the advantage of sparsity in highlighting faulty variables, the percentages of contributions of faulty variables (x_1, x_4, x_{18}, x_{44}) are compared for SCVA and CVA in Fig. 7. The accumulated contribution from faulty variables under SCVA takes a higher percentage of total contributions for most samples and is

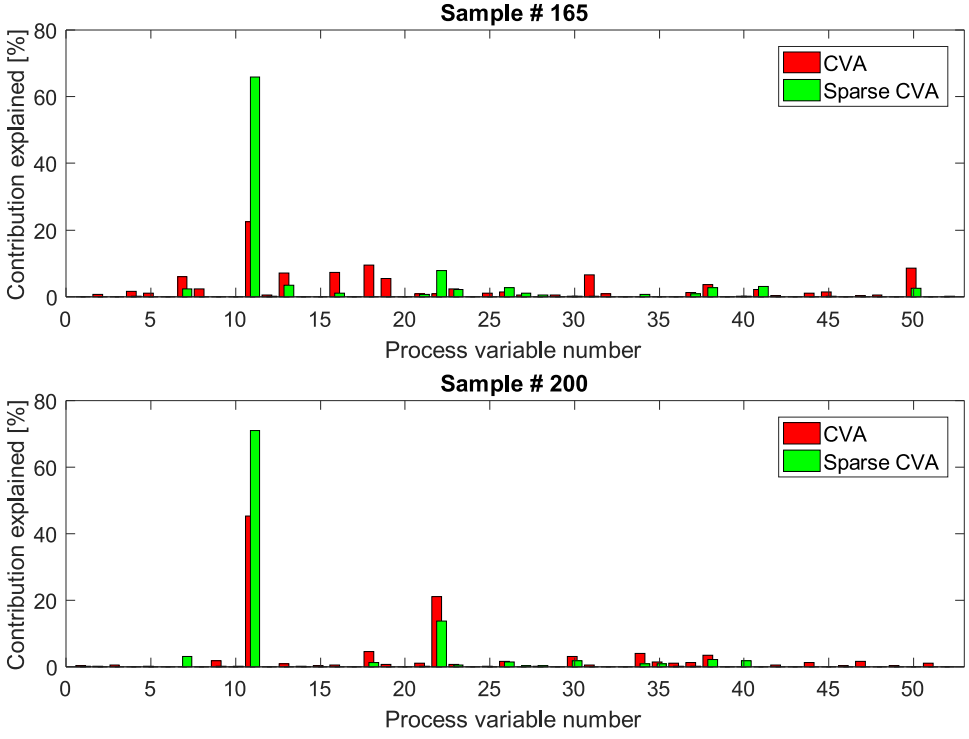

Fig. 8. Percentage of contributions of each variable for Fault 12 at samples 165 and 200.

much less noisy over time. The noisy total contributions from the CVA are mainly because CVA is not able to single out the tiny contributions from non-faulty variables and these small contributions are easily affected by the noise. However, the SCVA does not have this drawback due to the properties presented in previous sections. This observation highlights the advantage of SCVA in identifying faulty variables compared with traditional CVA.

The advantage of SCVA in fault identification is further illustrated by Fault 12 which is a random variation in the condenser cooling water inlet temperature. The condenser cooling water inlet temperature is not a directly measurable quantity, and the influence of Fault 12 is expected to be revealed by connected variables such as the condenser cooling water outlet temperature (x_{22}) and the product separator temperature (x_{11}). Similar to [24], the percentage of the contribution to T_d^2 of each variable is shown for sample 165 (5 samples after the fault takes place) and sample 200 in Fig. 8. At sample 165, the process variables x_{11} and x_{22} account for about 75% of the overall contributions for SCVA while accounting for less than 30% for CVA. Similarly, at sample 200 after which the control loop has deployed corrective actions, the process variables x_{11} and x_{22} explain about 85% of all contributions under SCVA and only 70% for CVA. Based on Figs. 7 and 8, SCVA intensifies the contributions from the faulty variables to be more distinct than the normal variables.

5. Conclusions and future work

This article presents a sparse canonical variate analysis approach for fault detection and identification. SCVA is preferred when the sample covariance matrices are close to singular in the case of collinear variables or small sample size. The sparsity parameter in SCVA trades off between sparsity and loss of information, which affects the fault detection performance. A recommended way of selecting the sparsity parameter is through cross-validation. Simulation results show that, with such a sparsity parameter, SCVA can achieve better fault detection performance than CVA for the Tennessee Eastman process. Moreover, SCVA preserves important variables in the canonical vectors, which improves the interpretability of sparse canonical vectors and uncovers important relationships among process variables. SCVA's sparse canonical vectors enable the determination of accumulated contributions on faulty variables so that they are more easily distinguished from normal variables. The results are verified in several TEP case studies. The proposed SCVA method can be extended to other processes that admit specific features, such as processes with multi-mode transitions [35]. Moreover, how to fully extract input- and output-relevant variations from residual space in SCVA is another direction of interest [36].

Acknowledgements

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) and by the Vanier Canada Graduate Scholarships (Vanier CGS). B. Jiang is grateful for the financial support from the National Natural Science Foundation of China (61603024). R.D. Braatz acknowledges the Edwin R. Gilliland Professorship.

Appendix A. Lower and Upper Bounds on c_1 for sparse CVA

The fact that c_1 is bounded below by 1 and above by $\sqrt{n_y l + n_u l}$ follows the fact from linear algebra that any vector $\boldsymbol{\alpha}$ of dimension $n_y l + n_u l$ must satisfy the bounds $\|\boldsymbol{\alpha}\|_2 \le \|\boldsymbol{\alpha}\|_1 \le \sqrt{n_y l + n_u l}\|\boldsymbol{\alpha}\|_2$. First consider violation of the stated lower bound $c_1 \ge 1$. For fixed $\boldsymbol{\beta}$, if $c_1 < 1$, then $\|\boldsymbol{\alpha}\|_1 \le c_1$ in (10) is a tighter constraint than $\|\boldsymbol{\alpha}\|_2^2 \le 1$ (since $\|\boldsymbol{\alpha}\|_2 \le \|\boldsymbol{\alpha}\|_1$, $\forall\boldsymbol{\alpha}$), the solution to (10) will not satisfy $\|\boldsymbol{\alpha}\|_2^2 \le 1$ as an equality, and the corresponding optimal $\boldsymbol{\alpha}$ cannot satisfy the definition of CVA. Now consider violation of the stated upper bound $c_1 \le \sqrt{n_y l + n_u l}$. If $c_1 > \sqrt{n_y l + n_u l}$, then $\|\boldsymbol{\alpha}\|_1 \le \sqrt{n_y l + n_u l}\,\|\boldsymbol{\alpha}\|_2 < c_1$ for any $\boldsymbol{\alpha}$ that meets $\|\boldsymbol{\alpha}\|_2^2 \le 1$, which implies that the l_1 bound constraint in (10) would not be active. In other words, the l_1 bound would be immaterial, the conventional CVA would be obtained, and sparsity would not be achieved.

References

[1] B. Jiang, D. Huang, X. Zhu, F. Yang, R.D. Braatz, Canonical variate analysis-based contributions for fault identification, J. Process Control 26 (1) (2015) 17–25.
[2] B. Jiang, X. Zhu, D. Huang, J.A. Paulson, R.D. Braatz, A combined canonical variate analysis and Fisher discriminant analysis (CVA-FDA) approach for fault diagnosis, Comput. Chem. Eng. 77 (9) (2015) 1–9.
[3] R.J. Treasure, U. Kruger, J.E. Cooper, Dynamic multivariate statistical process control using subspace identification, J. Process Control 14 (3) (2004) 279–292.
[4] J.E. Jackson, Multivariate quality control, Commun. Stat.-Theory Methods 14 (11) (1985) 2657–2688.
[5] R.B. Crosier, Multivariate generalizations of cumulative sum quality-control schemes, Technometrics 30 (3) (1988) 291–303.
[6] C.A. Lowry, W.H. Woodall, C.W. Champ, S.E. Rigdon, A multivariate exponentially weighted moving average control chart, Technometrics 34 (1) (1992) 46–53.
[7] J.V. Kresta, J.F. MacGregor, T.E. Marlin, Multivariate statistical monitoring of process operating performance, Can. J. Chem. Eng. 69 (1) (1991) 35–47.
[8] H. Lütkepohl, New Introduction to Multiple Time Series Analysis, second edition, Springer, Berlin Heidelberg, Germany, 2005.
[9] W. Ku, R.H. Storer, C. Georgakis, Disturbance detection and isolation by dynamic principal component analysis, Chemom. Intell. Lab. Syst. 30 (1) (1995) 179–196.
[10] L. Ljung, System Identification: Theory for the User, Prentice Hall, New Jersey, 1999.
[11] A. Negiz, A. Çinar, Statistical monitoring of multivariable dynamic processes with state-space models, AICHE J. 43 (8) (1997) 2002–2020.
[12] E.L. Russell, L.H. Chiang, R.D. Braatz, Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis, Chemom. Intell. Lab. Syst. 51 (1) (2000) 81–93.
[13] W.E. Larimore, Statistical optimality and canonical variate analysis system identification, Signal Process. 52 (2) (1996) 131–144.
[14] L.H. Chiang, E.L. Russell, R.D. Braatz, Fault Detection and Diagnosis in Industrial Systems, Springer Verlag, London, 2000.
[15] I. Wilms, C. Croux, Sparse canonical correlation analysis from a predictive point of view, Biom. J. 57 (7) (2015) 834–851.
[16] H.D. Vinod, Canonical ridge and econometrics of joint production, J. Econom. 4 (2) (1976) 147–166.
[17] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, J. Comput. Graph. Stat. 15 (2) (2006) 265–286.
[18] M. Journee, Y. Nesterov, P. Richtarik, R. Sepulchre, Generalized power method for sparse principal component analysis, J. Mach. Learn. Res. 11 (2) (2010) 517–553.
[19] A. d'Aspremont, L. El Ghaoui, M.I. Jordan, G.R.G. Lanckriet, A direct formulation for sparse PCA using semidefinite programming, SIAM Rev. 49 (3) (2007) 434–448.
[20] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, J. R. Stat. Soc. Ser. B Stat. Methodol. 72 (1) (2010) 3–25.
[21] E. Parkhomenko, D. Tritchler, J. Beyene, Sparse canonical correlation analysis with application to genomic data integration, Stat. Appl. Genet. Mol. Biol. 8 (1) (2009) 1–34.
[22] D.M. Witten, R. Tibshirani, T. Hastie, A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis, Biostatistics 10 (3) (2009) 515–534.
[23] S. Waaijenborg, P.C. Verselewel de, Witt Hamer, A.H. Zwinderman, Quantifying the association between gene expressions and DNA-markers by penalized canonical correlation analysis, Stat. Appl. Genet. Mol. Biol. 7 (1) (2008) 3.
[24] S. Gajjar, M. Kulahci, A. Palazoglu, Real-time fault detection and diagnosis using sparse principal component analysis, J. Process Control 67 (SI) (2018) 112–128.
[25] H. Gao, S. Gajjar, M. Kulahci, Q. Zhu, A. Palazoglu, Process knowledge discovery using sparse principal component analysis, Ind. Eng. Chem. Res. 55 (46) (2016) 12046–12059.
[26] S. Bao, L. Luo, J. Mao, D. Tang, Improved fault detection and diagnosis using sparse global-local preserving projections, J. Process Control 47 (1) (2016) 121–135.
[27] T. De Bie, N. Cristianini, R. Rosipal, Eigenproblems in pattern recognition, in: Handbook of Geometric Computing, Springer, 2005, pp. 129–167.

[28] W. Larimore, Canonical variate analysis in control and signal processing, Stat. Methods Control Signal Process. (1997) 83–120.
[29] H. Hotelling, Relations between two sets of variates, Biometrika 28 (3/4) (1936) 321–377.
[30] S. Dudoit, J. Fridlyand, T. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data, J. Am. Stat. Assoc. 97 (457) (2002) 77–87.
[31] L. Luo, S. Bao, J. Mao, D. Tang, Fault detection and diagnosis based on sparse PCA and two-level contribution plots, Ind. Eng. Chem. Res. 56 (1) (2016) 225–240.
[32] L. Xie, X. Lin, J. Zeng, Shrinking principal component analysis for enhanced process monitoring and fault isolation, Ind. Eng. Chem. Res. 52 (49) (2013) 17475–17486.
[33] X. Zhu, R.D. Braatz, Two-dimensional contribution map for fault identification, IEEE Control. Syst. 34 (5) (2014) 72–77.
[34] J. Liu, Fault diagnosis using contribution plots without smearing effect on non-faulty variables, J. Process Control 22 (9) (2012) 1609–1623.
[35] Y. Zhang, S. Li, Modeling and monitoring between-mode transition of multimodes processes, IEEE Trans. Ind. Inform. 9 (4) (2013) 2248–2255.
[36] Y. Zhang, Y. Fan, W. Du, Nonlinear process monitoring using regression and reconstruction method, Ieee Trans. Autom. Sci. Eng. 13 (3) (2016) 1343–1354.