Journal of Intelligent & Fuzzy Systems xx (20xx) x–xx
DOI:10.3233/JIFS-169550
IOS Press

Canonical variate analysis, probability approach and support vector regression for fault identification and failure time prediction

[a]*Faculty of Technology, De Montfort University, Leicester, UK*
[b]*School of Engineering, London South Bank University, London, UK*
[c]*Department of Rotating Equipment, Royal Dutch Shell, Hague, AN, The Netherlands*

Abstract. Reciprocating compressors are widely used in oil and gas industry for gas transport, lift and injection. Critical compressors that compress flammable gases and operate at high speeds are high priority equipment on maintenance improvement lists. Identifying the root causes of faults and estimating remaining usable time for reciprocating compressors could potentially reduce downtime and maintenance costs, and improve safety and availability. In this study, Canonical Variate Analysis (CVA), Cox Proportional Hazard (CPHM) and Support Vector Regression (SVR) models are employed to identify fault related variables and predict remaining usable time based on sensory data acquired from an operational industrial reciprocating compressor. 2-D contribution plots for CVA-based residual and state spaces were developed to identify variables that are closely related to compressor faults. Furthermore, a SVR model was used as a prognostic tool following training with failure rate vectors obtained from the CPHM and health indicators obtained from the CVA model. The trained SVR model was utilized to estimate the failure degradation rate and remaining useful life of the compressor. The results indicate that the proposed method can be effectively used in real industrial processes to perform fault diagnosis and prognosis.

Keywords: Condition monitoring, canonical variate analysis, cox proportional hazard model, support vector regression

1. Introduction

Modern industrial facilities such as natural-gas processing plants are becoming increasingly complex and large-scale as a result of increased mechanization and automation. The complexity of large-scale industrial facilities makes it difficult to build first-principle dynamic models for health monitoring and prognostics [9]. The existing condition monitoring approaches for industrial processes are typically derived from routinely collected system operating data. With the rapid growth and advancement in sensing and data acquisition technologies, long-term continuous measurements can be taken from different sensors mounted on machinery systems. However, using condition monitoring data for reliable faults diagnosis and prognosis remains a challenge for researchers and engineers.

A number of multivariate statistical techniques have been developed based on condition monitoring data for diagnostic and prognostic health monitoring, such as filtering based models [6], multivariate

*Corresponding author. David Mba, Faculty of Technology, De Montfort University, Leicester, LE1 9BH, UK. E-mail: david.mba@dmu.ac.uk.

time-series models [11] and neural networks [22]. Some of the key challenges in the implementation of these techniques are strongly correlated variables, high-dimensional data, changing operating conditions and inherent system uncertainty [4]. Recent developments of dimensionality reduction techniques have shown improvements in identifying faults from highly correlated process variables. Conventional dimensionality reduction methods are principal component analysis (PCA) [10], independent component analysis (ICA) [1] and partial least-squares analysis (PLSA) [21]. These basic multivariate methods have been proven to perform well under the assumption that process variables are time-independent. However, this assumption might not hold true for real industrial processes (especially chemical and petrochemical processes) because sensory signals affected by noises and disturbances often show strong correlation between the past and future sampling points [4]. Therefore, a few variants of the standard multivariate approaches [13, 20, 24] were developed later to solve the time-independency problem, making them more suitable for dynamic processes monitoring. Aside from approaches derived from PCA, ICA and PLSA, the canonical variable analysis (CVA) is a subspace method which takes serial correlations between different variables into account. Hence, is particularly suitable for dynamic process modelling [19]. The effectiveness of CVA has been verified by extensive simulation study [16, 19] and data captured from experimental test rigs [7]. However, the effectiveness of CVA in real complex industrial processes has not been fully studied.

Once a fault is detected in industrial processes, a fault identification tool is desired to find the variables that are most likely related to the specific fault (e.g. the candidate faulty variables). Contribution plots are one of the most popular tools for identifying the variables with the largest deviations when a fault occurs [26]. The traditional one-dimensional contribution maps can only be used to perform fault identification at one time instant, and is useful when the fault propagation is fast and localized. In comparison, 2-D contribution plots, which assemble the variations at multiple time instants, can clearly demonstrate the contributions of different process variables over the entire fault propagation process. In this investigation, 2-D contribution maps are applied to both the canonical residual and state space to perform faulty variable identification. The combination of the two types of statistics (residual and state space) can provide more insights into the fault than using a single statistic.

Typical condition monitoring procedures involve a prognostic step after the detection of a fault to estimate the failure time of the system. In this study, a combined CVA-CPHM-SVR method is proposed to perform fault prognostics based on both condition monitoring and lifetime data. CVA is utilized to transform the multidimensional data obtained from diverse sensors into a one-dimensional vector, which can be used to indicate the health condition of the compressor. The calculated health indicators are subsequently utilized together with CPHM and SVR to predict the failure time of the machine.

In medical research field, the Cox Proportional Hazard Model (CPHM) has been widely used for analyzing death rate or the probability of recurrence of a disease with censored survival data [5]. But its effectiveness in mechanical prognostic area has not been fully studied and only a limited number of publications have addressed its applicability for failure prediction of rotating machines [2, 3]. In this study, the CPHM model is utilized to estimate the failure degradation rate of the compressor using lifetime data. The degradation rate vectors obtained from the CPHM model are treated as input vectors and the health indicators derived from the CVA model are regarded as target vectors to train a SVR model. After training, the SVR model is utilized to make predictions of compressor degradation rate and failure time.

2. Methodology

2.1. CVA-based contributions for faulty variable identification

The objective of CVA is to find the maximum correlation between two sets of variables [9]. In order to generate two data matrices from the measured data $y_t \in \mathcal{R}^n$ (n indicates that there are n variables being recorded at each sampling time t), it was expanded at each sampling time by including p number of previous and f number of future samples to construct the past and future sample vectors $y_{p,t} \in \mathcal{R}^{np}$ and $y_{f,t} \in \mathcal{R}^{nf}$.

$$y_{p,t} = \begin{bmatrix} y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p} \end{bmatrix} \in \mathcal{R}^{np} \quad (1)$$

$$y_{f,t} = \begin{bmatrix} y_t \\ y_{t+1} \\ \vdots \\ y_{t+f-1} \end{bmatrix} \in \mathcal{R}^{nf} \tag{2}$$

To avoid the domination of variables with larger absolute values, the past and future sample vectors were then normalized to zero mean vectors $\widetilde{y_{p,t}}$ and $\widetilde{y_{p,t}}$, respectively. Then the vectors $\widetilde{y_{p,t}}$ and $\widetilde{y_{p,t}}$ at different sampling times were rearranged according to Equations (3) and (4) to produce the reshaped matrices $\hat{Y}_p$ and $\hat{Y}_f$:

$$\hat{Y}_p = \left[\hat{y}_{p,t+1},\ \hat{y}_{p,t+2}, \ldots,\ \hat{y}_{p,t+N}\right] \in \mathcal{R}^{np\times N} \tag{3}$$

$$\hat{Y}_f = \left[\hat{y}_{f,t+1},\ \hat{y}_{f,t+2}, \ldots,\ \hat{y}_{f,t+N}\right] \in \mathcal{R}^{nf\times N} \tag{4}$$

Where $N = l - p - f + 1$, and l represents the total number of samples for y_t.$\hat{Y}_p$ and $\hat{Y}_f$ are then processed by using the Cholesky decomposition to form a Hankel matrix $\mathcal{H}$ [18]. The purpose of using Cholesky is to form a new correlation matrix with reduced dimensionality such that the subsequent calculations could be conducted in a stable and fast manner. To find the linear combination that maximizes the correlation between the two sets of variables, the truncated Hankel matrix $\mathcal{H}$ is then decomposed by using Singular Value Decomposition (SVD):

$$\mathcal{H} = \sum\nolimits_{p,p}^{-1/2} \sum\nolimits_{p,f} \sum\nolimits_{f,f}^{-1/2} = U \sum V^T \tag{5}$$

Where $\Sigma_{p,p}$ and $\Sigma_{f,f}$ are the sample covariance matrices and $\Sigma_{p,f}$ denotes the cross-covariance matrix of $\hat{Y}_p$ and $\hat{Y}_f$.

If the order of the truncated

Hankel matrix $\mathcal{H}$ is d, then U, V and $\sum$ have the following form:

$$U = [u_1,\ u_2, \ldots,\ u_d] \in \mathcal{R}^{np\times d}$$

$$V = [v_1,\ v_2, \ldots,\ v_d] \in \mathcal{R}^{nf\times d}$$

$$\sum = \begin{bmatrix} d_1 & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_d \end{bmatrix} \in \mathcal{R}^{d\times d}$$

The columns of $U = [u_1,\ u_2, \ldots,\ u_d]$ and the columns of $V = [v_1,\ v_2, \ldots,\ v_d]$ are called the left-singular and right-singular vectors of $\mathcal{H}$, respectively. $\sum$ is a diagonal matrix, and its diagonal elements are called singular values, which depict the degree of correlation between the corresponding left-singular and right-singular vectors. The right-singular vectors in V corresponding to the largest r singular values were retained in the truncated matrix $V_r = [v_1,\ v_2, \ldots,\ v_r] \in \mathcal{R}^{np\times r}$. This matrix will be used later to perform dimension reduction on the measured data.

With the truncated matrix V_r, the np dimensional past vector $\hat{Y}_p \in \mathcal{R}^{np\times N}$ can be further converted into a reduced r-dimensional matrix $\Phi \in \mathcal{R}^{r\times N}$ (the columns of Φ are z_t, which are called state or canonical variates) by:

$$\Phi = [z_{t=1},\ z_{t=2}, \ldots,\ z_{t=N}] = J \cdot \hat{Y}_p \tag{6}$$

Similarly, the residual variates $\Psi \in \mathcal{R}^{np\times N}$ can be calculated according to Equation (7):

$$\Psi = [\varepsilon_{t=1},\ \varepsilon_{t=2}, \ldots,\ \varepsilon_{t=N}] = L \cdot \hat{Y}_p \tag{7}$$

where J and L are the projection matrices, and can be computed as: $J = V_r^T \sum_{p,p}^{-1/2} \in \mathcal{R}^{r\times np}$ and $L = V_e^T \sum_{p,p}^{-1/2} \in \mathcal{R}^{np\times np}$. Where V_r^T contains the first r columns of matrix V and V_e^T contains the $e = nf - r$ columns of V.

For a new observation y_t, the CVA-based state space contributions at time instant t can be computed from the state variates as:

$$\begin{aligned} c_t^{state} &= \left(J \cdot \hat{Y}_{p,t}\right)^T \left(J \cdot \hat{Y}_{p,t}\right) \\ &= \left(J \cdot \hat{Y}_{p,t}\right)^T \sum_{i=1}^{r} \left(\hat{Y}_{p,t} J_i^T\right)^T \\ &= \sum\nolimits_{i=1}^{r} \left(\hat{Y}_{p,t} J_i^T\right) \left(\hat{Y}_{p,t} J_i^T\right)^T \end{aligned} \tag{8}$$

Where $\hat{Y}_{p,t}$ denotes the column vector of $\hat{Y}_p$ at time instant t. J_i is the ith row of matrix J. Similarly, CVA-based residual space contributions at time instant t can be computed as:

$$\begin{aligned} c_t^{residual} &= \left(L \cdot \hat{Y}_{p,t}\right)^T \left(L \cdot \hat{Y}_{p,t}\right) \\ &= \left(L \cdot \hat{Y}_{p,t}\right)^T \sum_{i=1}^{np-r} \left(\hat{Y}_{p,t} L_i^T\right)^T \\ &= \sum_{i=1}^{np-r} \left(\hat{Y}_{p,t} L_i^T\right) \left(\hat{Y}_{p,t} L_i^T\right)^T \end{aligned} \tag{9}$$

The higher the contribution of a performance variable is, the larger the deviation of the specific variable from its normal value can be seen. Candidate faulty variables found in the canonical state space are related to large deviations of the system state present in healthy datasets. Whereas candidate faulty variables found in the canonical residual space are related to new system states generated during the monitoring process, which can no longer be fully described by the state space variates [12]. According to the literature [4], a limitation of CVA model is that the calculated contributions can be excessively sensitive because the inversion procedure of $\sum_{p,p}^{-1/2}$, which would result in incorrect identification of faulty variables. In order to alleviate this sensitivity, the combination of residual and state space contributions was adopted for the identification of variables most closely associated with the fault in this study, and this topic will be discussed in detail in Section 3.

2.2. CVA-based health monitoring

Aside from faulty variable identification, CVA is also a dimensionality reduction technique to monitor the machine operation by transferring the high-dimensional process data into one-dimensional health indicators. Condition monitoring data captured from the system operating under healthy conditions were used to calculate the threshold for normal operating limits. Abnormal operating conditions can be detected when the value of the health indicator exceeds the pre-set limits.

The canonical variates matrix Φ obtained from Equation (6) consists of valuable information that is needed to construct health indicators. The health indicator adopted in this study is the Hotelling statistics T^2 (introduced by Hotelling in 1936 [14]), which is the locus on the ellipse-like confidence region in the canonical variate space [15]. The Hotelling health indicator can be calculated as:

$$T_t^2 = \sum_{i=1}^{r} z_{t,i}^2 \tag{10}$$

Process data acquired during normal operating conditions were used to identify optimal threshold values of the Hotelling health indicator T_t^2. Since the Gaussian distribution doesn't hold true for non-linear processes, the actual probability density function of the health indicator was calculated by using a method named Kernel Density Estimation (KDE) [17]. Machine faults were considered every time when the health indicator exceeds the calculated threshold. The number of false detections was used in this study to determine the optimal number of retained state r, and the false detection was considered in two situations: (1) there is a violation of the Hotelling health indicator T_t^2 before the occurrence of fault; (2) the value of T_t^2 is smaller than the threshold determined by KDE after the occurrence of fault.

2.3. Cox proportional hazard model

Machinery fault degradation can be predicted by analyzing either condition monitoring measurements or historical lifetime data [25]. The CPHM, proposed by Cox [8], attempts to use both types of information for prognostic analysis of machinery fault degradation and failure times. A lifetime data set consists of failure times T of the machine under study, recorded either at failure time or before the final failure. In some cases, maintenance actions may be taken prior to failure to prevent a device or component from failing. Then these cases are considered as censoring since the actual failure time is unknown. In these cases, the recorded lifetime data is called censored data. The condition monitoring measurements used in CPHM can be any sensory signal that reflects the machine health condition.

CPHM assumes that the hazard rate or failure rate of a machine depends on two factors: the baseline hazard rate and the effects of covariates (condition measurements). Hence, the hazard rate of a machine at service time t can be written as:

$$h(t) = h_0(t)\, exp\left(\sum_{k=1}^{p} \beta_k Z_k\right) \tag{11}$$

Where $h_0(t)$ is called the baseline hazard function (It reflects the failure rate due to aging); $exp\left(\sum_{k=1}^{p} \beta_k Z_k\right)$ is the covariate function that describes how the covariates Z_k influence health degradation. The covariates are weighted through the regression parameters β_k. The estimation of the regression parameters is achieved by using a method called partial likelihood approach, which was proposed by Cox in 1972 [8]. According to Cox's theory, the partial likelihood of β_k can be written as:

$$L(\beta) = \prod_{i=1}^{n} \frac{exp\left(\sum_{k=1}^{p} \beta_k Z_{ik}(t_i)\right)}{\sum_{j \in R(t_i)} exp\left(\sum_{k=1}^{p} \beta_k Z_{jk}(t_j)\right)} \tag{12}$$

Then the optimal regression parameters can be estimated by maximizing the log likelihood of β_k:

$$LL(\beta) = \sum_{i=1}^{n}\sum_{k=1}^{p}\beta_k Z_{ik}(t_i) - \sum_{i=1}^{n} ln\left[\sum_{j\in R(t_i)} exp\left(\sum_{k=1}^{p}\beta_k Z_{jk}(t_j)\right)\right] \quad (13)$$

After model parameters are estimated, the hazard function can be calculated as:

$$\hat{h}_0(t_i;\hat{\beta}) = \frac{1}{\sum_{j\in R_{(t_i)}} exp\left(\sum_{h=1}^{p}\hat{\beta}_h Z_{jh}(t_j)\right)} \quad (14)$$

Then the cumulative hazard function and machine degradation rate can be approximated by formula (12) and (13), respectively:

$$\hat{H}(t) = \sum_{t_i\le t}\hat{h}(t_i;\hat{\beta}) \quad (15)$$

$$\hat{S}(t) = exp\left[-\hat{H}(t)\right] \quad (16)$$

2.4. Support vector regression

SVR is a supervised nonlinear regression approach. Application of the SVR model in the field of rotating machinery health monitoring and prognostics has been reported in [23, 27]. The target of SVR is to learn the dependency of an input vector $\{x_i\}_{i=1}^{N}$ on a target vector $\{y_i\}_{i=1}^{N}$ to make accurate forecast of y based on unseen values of x. When performing nonlinear regression, a kernel function is often chosen to map nonlinear inputs into a higher dimensional feature space, after which a minimum linear margin fit can be found in that space to perform linear regression. The form of the model is given as:

$$y = f(x,\ w) = \sum_{i=1}^{N} w_i K(x,\ x_i) + b \quad (17)$$

where $w = (w_1,\ w_2, \ldots,\ w_N)^T$ is a weight vector, which elucidates the links between the high dimensional space and the target output; and $K(x, x_i)$ denotes the kernel function, and b denotes the bias.

A SVR model is first built based on the health indicators generated by CVA and the degradation rates obtained from CPHM. Then the trained SVR model is employed to predict degradation rate and failure time of the compressor given unseen input health indicators. The flowchart of the combined CVA-CPHM-SVR prognostic method is shown in Fig. 1.

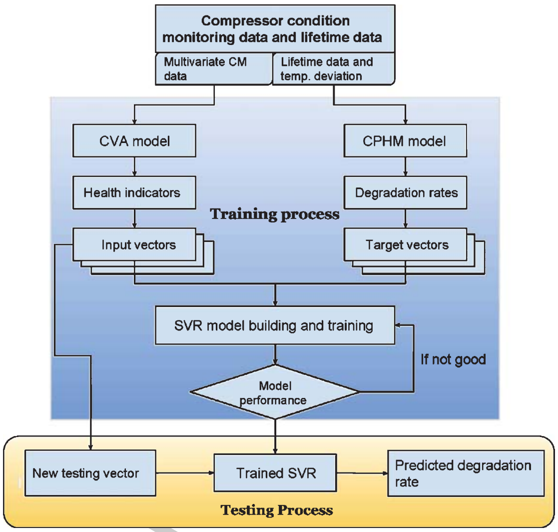

Fig. 1. Schematic diagram of the proposed prognostic method.

3. Validation using reciprocating compressor condition monitoring data

3.1. Data acquisition

Reciprocating compressors are widely used in oil and gas industry for gas transport, lift and injection. They typically operate under high rotating speed, high pressure and high load conditions, and are therefore subject to performance degradations. These machines are highly automated with various sensors being mounted all over the system, and signals from different sensors can be stored and accessed through an e-maintenance system. The data used in this study were gathered from a two-stage, four-cylinder, double-acting reciprocating compressor used in a refinery in Europe.

The compressor experienced twelve valve failures at cylinder 4 from July 2013 to December 2014. Machine inspections revealed that the failure mode under study was valve leakage caused by broken valve plate. The failed valves were either the head end or the crank end discharge valve. A total of 12 fault cases were obtained from the site engineer and each sample was a multivariate time series consisting of 39

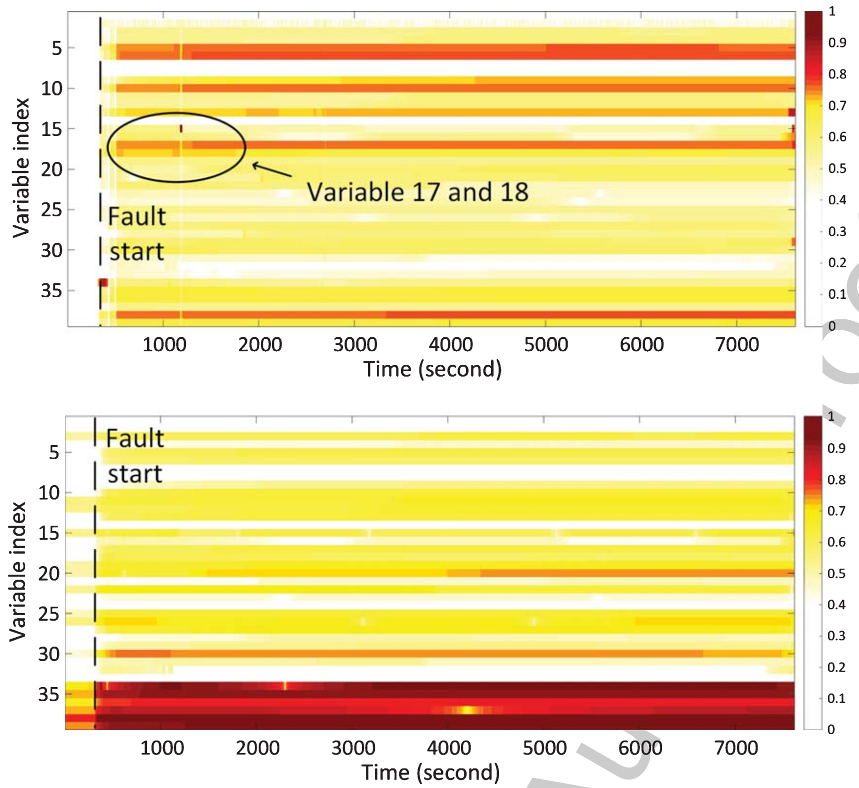

Fig. 2. CVA-based contribution plots for faulty variable identification in fault case 3: (1) faulty variables identified in residual space (upper); (2) faulty variables identified in state space (lower). Contributions are normalized to a range of 0 to 1.

variables. The sampling rate was 1 Hz and the failure degradation duration for each sample was different.

3.2. *CVA-based contributions for faulty variable identification*

Once a fault occurs in industrial heavy-duty compressors, it is important to identify which components are most likely associated with the root-cause of the malfunction. Contribution plot analysis [4] is one of the most popular tool for identifying "fault related" variables in multidimensional statistical analysis. In this section, CVA-based state space and residual space contributions were used to identify candidate faulty variables for the compressor under study. The contributions of different process variables in fault case 3 were depicted in Fig. 2 using color map with variable number being the vertical axis and sampling time being the horizontal axis. As stated previously, the root cause of the fault was discharge valve failure in cylinder 4, meaning that the most fault related variables were variable 17 and 18 (highlighted in bold in Table 1). As shown in Fig. 2, the residual space 2-D

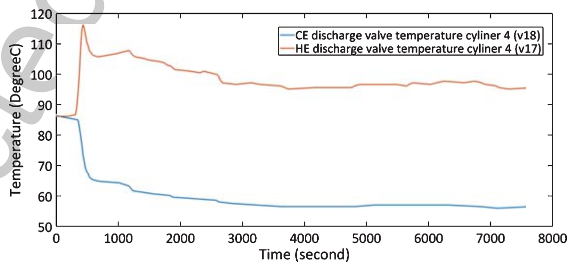

Fig. 3. Trends of the HE and CE discharge valve temperature in cylinder 4 for fault case 3.

map indicates high contributions of both variable 17 and 18 during the early stage of fault case 3. Then the contribution of variable 18 dropped to a lower level after around the 1500th sampling point, whereas variable 17 continued to show high contributions until the end of the sampling period. By looking closely at the trends of variable 17 and 18 (see Fig. 3), it was found that with the compressor controller applied to the system, variable 18 stabilized to its normal operating range after about the 1500th sample. However, due to the malfunction of HE discharge valve

Table 1
Identified candidate faulty variables for all fault cases

Variable No.	Variable name	F1	F2	F3	F4	F5	F6	F8	F9	F 10	F 11	F 12	F 13
1	Speed						B						
2	Actual total flow												
3	HE suction valve temperature cylinder 1	R					B		R			R	
4	CE suction valve temperature cylinder 1	R					B		R				
5	HE discharge valve temperature cylinder 1	R	R	B			B	B	B	Y		Y	R
6	CE discharge valve temperature cylinder 1	Y	Y	B	Y		B	B	B	B		B	Y
7	HE suction valve temperature cylinder 2						B						
8	CE suction valve temperature cylinder 2						B						
9	HE discharge valve temperature cylinder 2	Y	R		B		B		B			B	
10	CE discharge valve temperature cylinder 2	Y	Y	B	B	R	B	B	B	B		Y	
11	HE suction valve temperature cylinder 3	R	B	B	R		B					R	
12	CE suction valve temperature cylinder 3		B				Y						
13	HE discharge valve temperature cylinder 3			B			Y	B	B	B		B	Y
14	CE discharge valve temperature cylinder 3												
15	HE suction valve temperature cylinder 4					B							
16	CE suction valve temperature cylinder 4		R			B	B						
17	*HE discharge valve temperature cylinder 4*			B	B	B	B	B	B	B		B	
18	*CE discharge valve temperature cylinder 4*	B	R			B	B	B	B	B		B	
19	Main bearing temperature 1	R			R	Y	B			R			
20	Main bearing temperature 2	R		R	R	B	Y			R		R	R
21	Main bearing temperature 3	R			R	R	Y	R				R	
22	Main bearing temperature 4				R		Y		R				
23	Vent flow cylinder 1						Y		R				
24	Vent flow cylinder 2					R	Y		R				
25	Vent flow cylinder 3		Y				Y						
26	Vent flow cylinder 4			R			B			R		R	
27	Rod drop cylinder 1					R							
28	Rod drop cylinder 2							R					
29	Rod drop cylinder 3		B						B	R			
30	Rod drop cylinder 4			R									
31	Vibration crosshead 1										R	B	
32	Vibration crosshead 2		B										
33	Vibration crosshead 3												Y
34	Vibration crosshead 4			R				R	B			R	
35	Lube oil supply pressure	R	Y	R	Y	R	Y	Y	R	R	Y	R	Y
36	Lube oil reservoir level	R	R	R		R		Y	R	R	B		B
37	Lube oil supply temperature	Y	Y	R	Y	R	Y	R	R	R	Y		Y
38	Lube oil filter DP	R	Y	Y	Y	R	Y	Y	Y	R	Y		Y
39	Lube counter		B	R	Y	R	Y						

Note:
R (red): Candidate faulty variables identified in the state space
B (blue): Candidate faulty variables identified in the residual space
Y (yellow): Candidate faulty variables identified in both state space and residual space

in cylinder 4, large deviations from normal operating conditions were observed in variable 17 until the end of the sampling period. Therefore, variable 17 rather than variable 18 was considered as a candidate faulty variable in this case.

It is worth noting that in addition to variable 17 and 18, several other faulty variables were revealed by the residual and state space contributions. The reason these variables have large contributions is that the fault has propagated from cylinder 4 into other components, resulting in loss of performance of the entire compressor.

The identified candidate faulty variables for all fault cases are summarized in Table 1. Collectively,

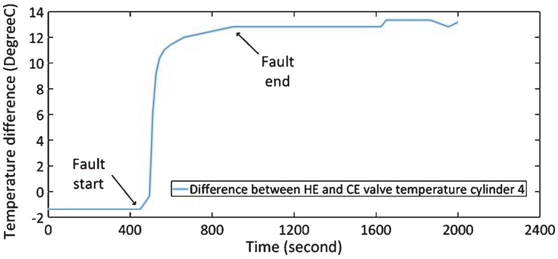

Fig. 4. Difference between CE and HE discharge temperature in cylinder 4 – failure sample No. 2.

CVA-based contributions are very effective at identifying the root cause of the compressor fault as the CE/HE discharge valve temperature in cylinder 4 has been successfully reported as a faulty variable in most cases. Collectively the identified candidate faulty variables would provide valuable information to a site engineer as to the fundamental cause of the fault. In addition, it was found that the root cause was more often linked to faulty variables identified in the residual space rather than in the state space. This demonstrates the necessity of combining residual and state space contributions for fault identification as utilizing merely the state space information can lead to wrong decision making.

3.3. Determination of fault start time fault end time

Since the failure mode under study is head end/crank end valve damage took place in cylinder 4, the method employed to determine the fault start and end time, as suggested by the site engineers, is to look at the difference between crank end (CE) discharge temperature and head end (HE) discharge temperature in cylinder 4. To be specific, during healthy operating conditions and after the failure point, as shown in Fig. 4, the temperature difference between CE and HE is relatively constant. However, the temperature difference grows continuously once the valve fault occurs.

As shown in Fig. 4, the fault start time for fault case 2 was identified when the value of temperature difference starts to increase, whereas the fault end time was identified when the temperature difference stabilized at its new steady state value. The degradation duration for all failure cases can be found in Table 2.

3.4. CVA model building

A CVA model was firstly built in order to transform the multivariate condition monitoring data into a one-dimensional health indicator. This process can be considered as a data fusion and dimensionality reduction procedure as it incorporates the information from all the measured 39 variables to generate a health indicator which can reflect the health condition of the system. For each fault case, a normal operating dataset was used to train the CVA algorithm to obtain the normal operating limits of T_t^2, and a deteriorating dataset was used to construct a health indicator.

In order to build a CVA model as described in Equations (1 to 7), three tuning parameters need to be determined, namely, the number of time lags p and f, and the number of dimensions retained r According to the literature [17], the number of time lags p and f were determined by calculating the autocorrelation function of the root summed squares of all variables against a confidence bound of $\pm 5\%$. The autocorrelation function indicates how long the measured time series is correlated with itself, and thus can be used to determine the maximum number of significant lags. As shown in Fig. 5, the sample autocorrelation analysis of the training data demonstrates that the maximum number of significant lags was 25. Therefore, the number of time lags p and f were set to 25 in this study.

Table 2
Degradation duration for all failure cases

Sample No.	Degradation Length (s)
6	171
11	191
3	231
1	371
13	381
10	391
5	401
8	441
2	451
4	501
12	601
9	641

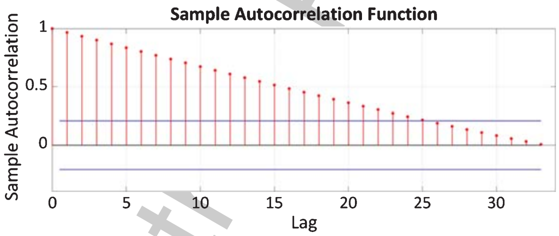

Fig. 5. Autocorrelation of the root summed squares of all variables in training dataset.

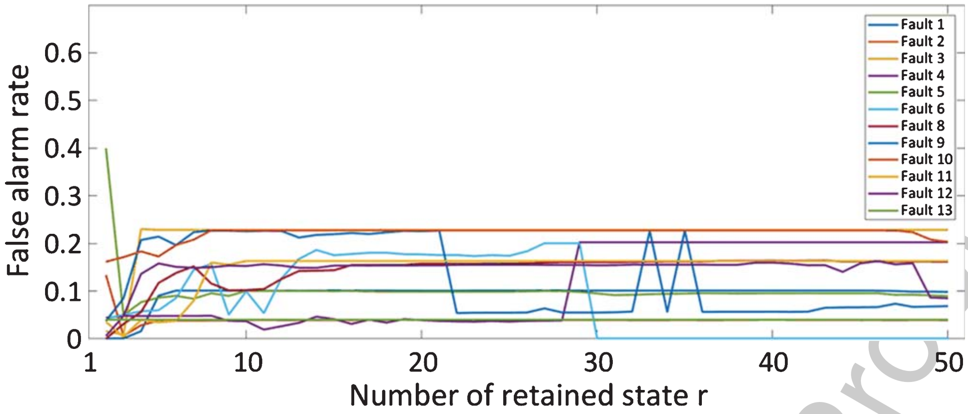

Fig. 6. False alarm rate of all fault cases with different values of r.

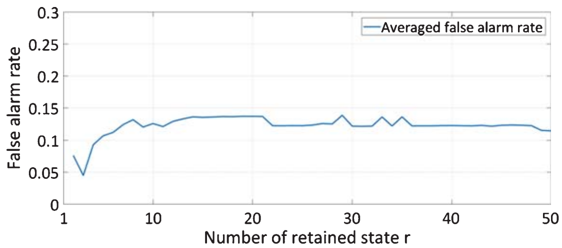

Fig. 7. Averaged false alarm rate with different values of r.

In order to determine the optimal number of r, CVA was implemented to perform fault detection for all 12 fault cases using different values of r. The false alarm rate versus the number of retained states for all fault cases were depicted in Fig. 6. False alarm rate in this study was calculated by dividing the number of false detections by the length of the testing dataset. Then the calculated false alarm rates were averaged with the purpose of selecting the optimal value of r that minimizes the false alarm rate for all fault cases. $r = 3$ was finally adopted according to the results shown in Fig. 7.

As discussed previously, the fault start and end times in this study were determined by looking at the difference between CE and HE discharge temperature in cylinder 4. The health indicators generated by the trained CVA model were further truncated according to the fault duration of specific fault cases. Figure 8 depicts the truncated health indicators for all 12 failure cases. They will be used hereafter as target vectors for SVR training.

3.5. *CPHM model building*

In order to build a CPHM model, lifetime data of 12 fault cases were used to estimate the baseline hazard function. In addition, the difference between CE and HE discharge temperature in cylinder 4 was assumed as a covariate and the regression parameter β_k was calculated as per Equations (12 and 13)

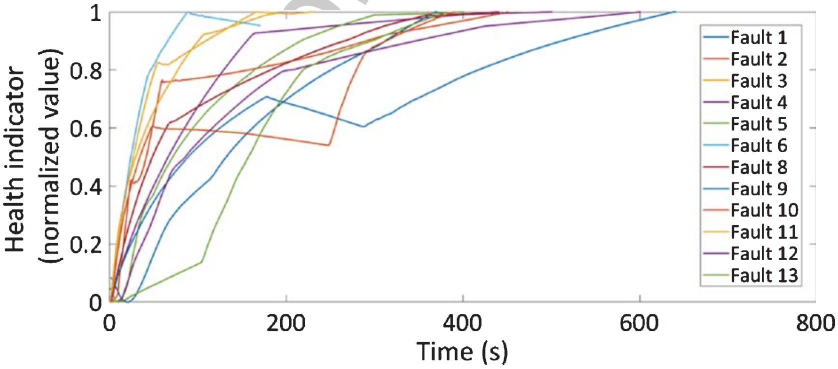

Fig. 8. Truncated health indicators of all fault cases.

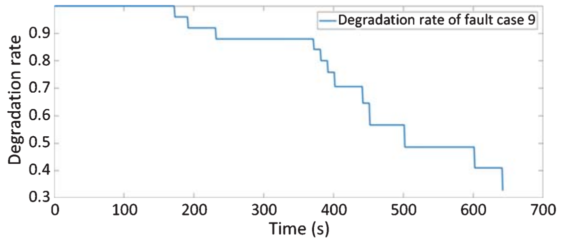

Fig. 9. Hazard rate of failure sample no. 9.

for each failure case. For example, Fig. 9 shows the calculated degradation rate of fault case 9.

3.6. SVR model building and testing

In this section, health indicators and failure rate vectors obtained previously were used to train a SVR model. Then the trained SVR was employed as a prognostic method to predict the failure degradation of individual failure case. To build a SVR model, we utilized a Radial Basis Function (RBF) kernel function to map input vectors into the high-dimensional feature space. The RBF kernel parameter γ and the soft margin parameter C were determined using grid search [28] together with 5-fold cross validation. For grid search, parameter γ and C take the following values:

The health indicator and degradation rate vector of fault case no. 10 were firstly utilized to train a SVR model. The optimal parameters determined by grid search were 1024 and 64 for γ and C, respectively. They were determined by searching for the minimum Root-Mean-Squared Error (RMSE) between the actual degradation rate and the estimated degradation rate for each combination of γ and C candidates (as shown in Fig. 10). Moreover, the health indicator of fault case no. 13 was used as an input vector to test the performance of the trained SVR model. The predicted degradation rate of fault no. 13 is depicted in Fig. 11. It can be observed that the predicted failure time is 381 s.

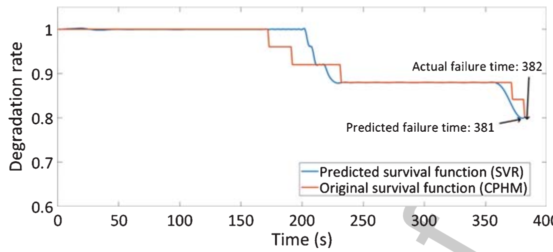

Fig. 11. SVR prediction for fault case no. 13.

$$\gamma = 2^{\{-10,-9,-8,\ldots,\ 10\}}$$

$$C = 2^{\{-10,-9,-8,\ldots,\ 10\}}$$

In order to fully capture the dynamics of the compressor, a SVR model was further trained by 8 fault cases (F1, F13, F10, F5, F8, F4 and F12). The input vectors used to perform the training were obtained using the CVA method. In addition, the target vectors were acquired by an estimation of the degradation rate by means of CPHM. The optimal value of γ and C was 128 and 256 respectively according to the results of grid search. Figure 12 depicts the RMSE between the actual and the estimated target vectors for each combination of γ and C candidates. The trained SVR model was utilized to predict the hazard rate of fault case no. 2, and the predicted result is shown in Fig. 13. The predicted failure time is 449 s while the actual failure happens at 452 s.

The performance of the prognostic model can be assessed using the following metrics, namely Accuracy, root mean squared error (RMSE), mean absolute error (MAE) and Pearson's correlation coefficient (R). Formulae of the above metrics are listed as follows:

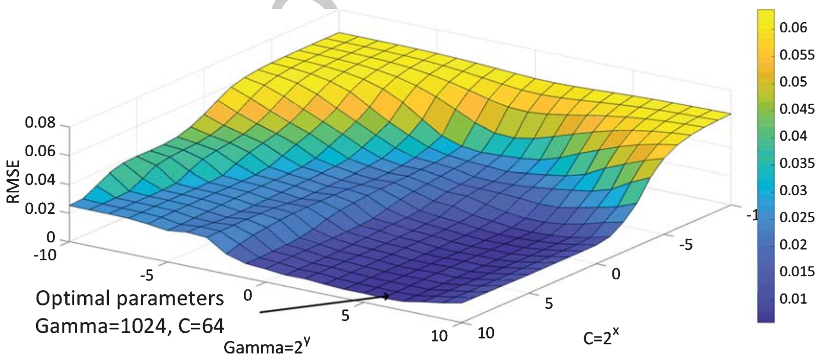

Fig. 10. RMSE for various values of γ and C model parameters.

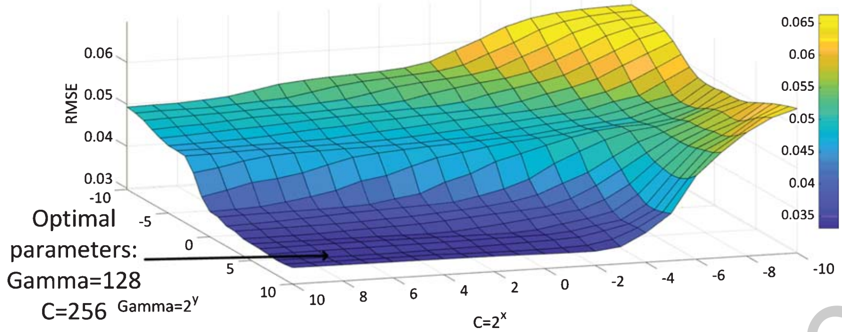

Fig. 12. RMSE for various values of γ and C model parameters (using f1, f13, f10, f5, f8, f4, and f12 for training).

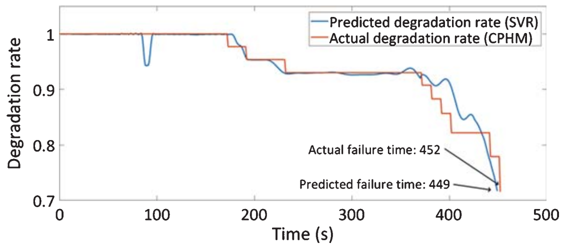

Fig. 13. Predicted failure rate of sample no. 2.

$$Accuracy = \left(1 - \frac{T_{actual} - T_{predicted}}{T_{actual}}\right) \times 100\% \tag{18}$$

$$RMSE = \left[\sum_{i=1}^{N} \left(S(t)_{actual,i} - S(t)_{predicted,i}\right)^2 / N\right]^{1/2} \tag{19}$$

$$MAE = \frac{1}{N}\sum_{i=1}^{N} \left|S(t)_{actual,i} - S(t)_{predicted,i}\right| \tag{20}$$

$$R = \frac{\sum_{i=1}^{N} \left(S(t)_{act,i} - \overline{S(t)_{act}}\right)\left(S(t)_{pre,i} - \overline{S(t)_{pre}}\right)}{\sqrt{\sum_{i=1}^{N} \left(S(t)_{act,i} - \overline{S(t)_{act}}\right)^2}\sqrt{\sum_{i=1}^{N} \left(S(t)_{pre,i} - \overline{S(t)_{pre}}\right)^2}} \tag{21}$$

A higher value of Accuracy indicates a better the prediction. Meanwhile, the higher the value of RMSE/MAE is, the lower the prediction accuracy is. A high Pearson's correlation coefficient means a high accordance between the actual and predicted degradation rate. The performance of the predictive model, based on the proposed four metrics, is summarized in Table 3. The predicted degradation rate of fault case no. 2 seems overestimated between 370 s and 430 s and underestimated between 431 s to 449 s, yielding a relatively high MAE value. But the accuracy is 99.33%, which is admissible for constructing the prognostic model.

Table 3
Model performance based on four statistical indexes

Sample No.	Accuracy	RMSE	MAE	R
13	99.74%	0.02	0.0082	0.9485
2	99.33%	0.0076	0.0482	0.933

4. Conclusion

In this study, condition monitoring data acquired from an operational industrial reciprocating compressor have been used to test the capabilities of CVA for fault identification. In addition, CVA combined with CPHM and SVR were applied for the first time to perform prognostics based on condition monitoring and lifetime data. 2-D contribution plots based on the variations in the residual and state spaces were utilized to identify candidate faulty variables for compressor faults. It was found that the fundamental causes are more likely to be related to the residual space. Furthermore, CPHM was utilized to calculate the fault degradation rate based on lifetime data obtained from the compressor, and the calculated degradation vectors were regarded as the target vectors for training a SVR model. Grid search and 5-fold cross validation were used to determine the optimal SVR model parameters during the training process. Finally, the trained SVR was employed to predict degradation rate and failure time of the compressor. Four metrics were utilized to evaluate the accuracy of the proposed scheme. The results illustrate that the prognostic performances were satisfied.

Although, the results of this study clearly show the superior performance of the proposed methods for fault identification and failure prediction, some aspects require further investigation are listed as follows. Firstly, apart from CE/HE discharge valve temperature in cylinder 4, several other faulty variables were reported by both the residual and state space contributions. A consideration for future work is to alleviate the smearing effect and reduce the number of reported faulty variables, thereby allowing for more accurate fault identification. Secondly, due to the approximative nature of hazard function, the degradation vectors used in this investigation are stair functions with jumps at failure times. Thus, a degradation curve might not truly reflect the deterioration process when the number of historical failures is small, which would lead to inaccurate failure time prediction. Hence, techniques should be developed to calculate machine degradation rates accurately regardless of the scarcity of lifetime data.

References

[1] A. Hyvärinen, J. Karhunen and E. Oja, *Independent Component Analysis*, Wiley, New York, 2004.

[2] A.K.S. Jardine, D. Banjevic, M. Wiseman, S. Buck and T. Joseph, Optimizing a mine haul truck wheel motors' condition monitoring program use of proportional hazards modeling, *Journal of Quality in Maintenance Engineering* **7** (2001), 286–302.

[3] A.K.S. Jardine, P.M. Anderson and D.S. Mann, Application of the Weibull proportional hazards model to aircraft and marine engine failure data, *Quality and Reliability Engineering International* **3** (1987), 77–82.

[4] B.B. Jiang, D.X. Huang, X.X. Zhu, F. Yang and R.D. Braatz, Canonical variate analysis-based contributions for fault identification, *Journal of Process Control* **26** (2015), 17–25.

[5] B. Zupan, J. Demsar, M.W. Kattan, J.R. Beck and I. Bratko, Machine learning for survival analysis: A case study on recurrence of prostate cancer, *Artificial Intelligence in Medicine* **20** (2000), 59–75.

[6] C.J. Guerra and J.R. Kolodziej, A data-driven approach for condition monitoring of reciprocating compressor valves, *Journal of Engineering for Gas Turbines and Power* **136** (2014), 041601.

[7] C.R. Cárcel, Y. Cao and D. Mba, A benchmark of canonical variate analysis for fault detection and diagnosis, *UKACC International Conference on Control (CONTROL), IEEE*, Loughborough, UK, 2014, pp. 425–431.

[8] D.R. Cox, Regression models and life-tables, in: Breakthroughs in Statistics, Springer, New York, 1992, pp. 527–541.

[9] E.L. Russell, L.H. Chiang and R.D. Braatz, Fault detection in industrial processes using canonical variate analysis and dynamic principal component analysis, *Chemometrics and Intelligent Laboratory Systems* **51** (2000), 81–93.

[10] F. Harrou, M.N. Nounou, H.N. Nounou and M. Madakyaru, Statistical fault detection using PCA-based GLR hypothesis testing, *Journal of Loss Prevention in the Process Industries* **26** (2013), 129–139.

[11] F. Serdio, E. Lughofer, K. Pichler, T. Buchegger, M. Pichler and H. Efendic, Fault detection in multi-sensor networks based on multivariate time-series models and orthogonal transformations, *Information Fusion* **20** (2014), 272–291.

[12] G. Li, S.J. Qin and T. Yuan, Data-driven root cause diagnosis of faults in process industries, *Chemometrics and Intelligent Laboratory Systems* **159** (2016), 1–11.

[13] G. Stefatos and A.B. Hamza, Dynamic independent component analysis approach for fault detection and diagnosis, *Expert Systems with Applications* **37** (2010), 8606–8617.

[14] H. Hotelling, Relations between two sets of variates, *Biometrika* **28** (1936), 321–377.

[15] J.E. Jackson, Quality control methods for several related variables, *Technometrics* **1** (1959), 359–377.

[16] L.Z. Huang, Y.P. Cao, X.M. Tian and X.G. Deng, A nonlinear quality-relevant process monitoring method with kernel input-output canonical variate analysis, *IFAC-PapersOnLine* **48** (2015), 611–616.

[17] P.-E.P. Odiowei and Y. Cao, Nonlinear dynamic process monitoring using canonical variate analysis and kernel density estimations, *IEEE Transactions on Industrial Informatics* **6** (2010), 36–45.

[18] R.T. Samuel and Y. Cao, Kernel canonical variate analysis for nonlinear dynamic process monitoring, *IFAC-PapersOnLine* **48** (2015), 605–610.

[19] S. Stubbs, J. Zhang and J. Morris, Fault detection in dynamic processes using a simplified monitoring-specific CVA state space modelling approach, *Computers & Chemical Engineering* **41** (2012), 77–87.

[20] S. Yin, X.P. Zhu and O. Kaynak, Improved PLS focused on key-performance-indicator-related fault diagnosis, *IEEE Transactions on Industrial Electronics* **62** (2015), 1651–1658.

[21] U. Kruger and G. Dimitriadis, Diagnosis of process faults in chemical systems using a local partial least squares approach, *AIChE Journal* **54** (2008), 2581–2596.

[22] V.T. Tran, F. AlThobiani and A. Ball, An approach to fault diagnosis of reciprocating compressor valves using Teager–Kaiser energy operator and deep belief networks, *Expert Systems with Applications* **41** (2014), 4113–4122.

[23] W. Caesarendra, A. Widodo, P.H. Thom, B.-S. Yang and J.D. Setiawan, Combined probability approach and indirect data-driven method for bearing degradation prognostics, *IEEE Transactions on Reliability* **60** (2011), 14–20.

[24] W.H. Li and S.J. Qin, Consistent dynamic PCA based on errors-in-variables subspace identification, *Journal of Process Control* **11** (2001), 661–678.

[25] X.C. Li, F. Duan, D. Mba and I. Bennett, Multidimensional prognostics for rotating machinery: A review, *Advances in Mechanical Engineering* **9** (2017), 1687814016685004.

[26] X.X. Zhu and R.D. Braatz, Two-dimensional contribution map for fault identification [Focus on Education], *IEEE Control Systems* **34** (2014), 72–77.
[27] Y. Qian and R. Yan, Remaining useful life prediction of rolling bearings using an enhanced particle filter, *IEEE Transactions on Instrumentation and Measurement* **64** (2015), 2696–2707.
[28] Z.L. Liu, M.J. Zuo and H.B. Xu, Parameter selection for Gaussian radial basis function in support vector machine classification, *International Conference on Quality, Reliability, Risk, Maintenance, and Safety Engineering, IEEE*, Chengdu, China 2012, pp. 576–581.