

# Dynamic fault detection and diagnosis for alkaline water electrolyzer with variational Bayesian Sparse principal component analysis

Qi Zhang<sup>a</sup>, Weihua Xu<sup>a</sup>, Lei Xie<sup>a,\*</sup>, Hongye Su<sup>a</sup>

<sup>a</sup>State Key Laboratory of Industrial Control Technology, Zhejiang University, 310027 Hangzhou, China

---

## Abstract

Electrolytic hydrogen production serves as not only a vital source of green hydrogen but also a key strategy for addressing renewable energy consumption challenges. For the safe production of hydrogen through alkaline water electrolyzer (AWE), dependable process monitoring technology is essential. However, random noise can easily contaminate the AWE process data collected in industrial settings, presenting new challenges for monitoring methods. In this study, we develop the variational Bayesian sparse principal component analysis (VBSPCA) method for process monitoring. VBSPCA methods based on Gaussian prior and Laplace prior are derived to obtain the sparsity of the projection matrix, which corresponds to  $\ell_2$  regularization and  $\ell_1$  regularization, respectively. The correlation of dynamic latent variables is then analyzed by sparse autoregression and fault variables are diagnosed by fault reconstruction. The effectiveness of the method is verified by an industrial hydrogen production process, and the test results demonstrated that both Gaussian prior and Laplace prior based VBSPCA can effectively detect and diagnose critical faults in AWEs.

*Keywords:*

Alkaline water electrolyzer, sparse principle component analysis, fault diagnosis, sparse Bayesian learning, variational Bayesian

---

## 1. Introduction

As the global energy transition accelerates, Alkaline water electrolyzer (AWE) for hydrogen production has become the most widely used hydrogen production technology. As industrial hydrogen production grows in scale, the safety of hydrogen production has attracted more attention. The process of producing hydrogen involves multiple scientific disciplines, and the system is particularly complex and difficult to control. With the widespread adoption of Distributed control systems (DCS) and sensor technology in industrial settings, it is possible to collect and store large amounts of data from the AWE process. Moreover, data mining and database technologies offer

---

\*Corresponding author

strong technical support for developing and applying data-driven modeling approaches in industrial AWE processes[1]. Hence, by extracting and analyzing information from process data, process monitoring models can be built for industrial applications to ensure safe operation of industrial water electrolyzer processes[2, 3].

Latent variable models such as Principal component analysis (PCA) and Partial least squares (PLS) enable the extraction of key information from the process to build efficient monitoring systems by projecting high-dimensional process data into a low-dimensional space[4]. However, both PCA and PLS under the assumption that the data involved in the process exhibit no temporal correlation. In fact, there is significant dynamic correlation in the process data. Therefore, neither of these methods exploits the dynamic correlation between the data. There are three types of Dynamic latent variable (DLV) methods that have been most extensively studied[5]. The most conventional approach is to construct time lag matrices, such as the dynamic PCA method, but this approach significantly reduces the interpretability of faults[6, 7]. In addition, dynamic relationships can be modelled explicitly by state space models, such as variational Bayesian state space models[8]. There are also methods such as dynamic inner PCA [9] and dynamic inner canonical correlation analysis (DiCCA) [10] which focus more on the dynamic relationships between latent variables. However, the disadvantage of the latent variable models is the high requirement for data quality[11]. The latent variable models look for the orthogonal direction with the largest data variance to separate noise and data as much as possible[12]. Given that industrial process data are easily contaminated by noise, latent variable models are sensitive to noise and outliers, which reduces the reliability of monitoring.

Sparse methods have a strong noise reduction capability that removes noise from the data while retaining critical information[13, 14]. This is especially important for fault detection in noisy industrial environments. Traditionally, sparsity is obtained by utilizing regularization in the loss function.  $\ell_1$  regularization and  $\ell_2$  regularization are the most common methods, corresponding to Lasso regression and Ridge regression, respectively[15]. The core of Lasso regression is that it uses the sum of the absolute values of the weight coefficients as the penalty term, which makes the loss function of Lasso regression discontinuously derivable[16]. The selection of the regularization factor is usually difficult, and it is empirically selected, resulting in non-robustness to changes in the algorithmic environment. Bayesian sparsity treats regularization as a Bayesian framework for prior knowledge[17, 18]. In this approach, the parameter estimates are updated with the posterior distributions, treating the regularisation terms as prior distributions of the model parameters[19]. In addition, sparse Bayesian learning (SBL) has become a hot topic of research because it improves the interpretability and generalisability of the model by taking parameter uncertainty into

account[20, 21]. Expectation maximisation (EM), which reduces computational cost and handles missing values, is a common approach to implementing sparse Bayesian methods[22, 23]. However, EM can only provide point estimates of parameters that may fall into local optima or overfitting, and requires strict assumptions about the posterior distributions of the hidden variables. Therefore, the full Bayesian uncertainty of the model cannot be captured by the EM method[24, 25]. A more advantageous approach involves utilizing variational Bayesian methods, which approximate the complete posterior distribution of parameters and hidden variables, thereby circumventing the constraints associated with the EM algorithm[26, 27].

The main objective of this study is to develop a process monitoring technique for the AWE hydrogen production process that can effectively handle noise. To achieve this, we have derived variational Bayesian principal component analysis (VBSPCA) methods for process monitoring based on Gaussian and Laplace priors, respectively, inspired by Bayesian sparsity. The introduction of  $\ell_2$  regularization is equivalent to introducing a Gaussian prior, while the introduction of  $\ell_1$  regularization is equivalent to introducing a Laplace prior. By employing matrices with a shared sparse structure and independent Gaussian priors for low-rank decomposition, Gaussian prior-based VBSPCA is capable of accurately modeling and estimating noise, thereby mitigating its effects. The Laplace prior-based VBSPCA designs a two-level hierarchical decomposition of the Laplace distribution to construct the conjugate structure for Bayesian posterior inference. In addition, the parameters of VBSPCA are estimated by variational Bayesian inference, which not only quantifies the model uncertainty, but also makes the algorithm easy to handle. Finally, to make full use of the structurally relevant information of the variables, we incorporate the vector autoregression (VAR) with  $\ell_1$  regularization to better extract dynamic features from the latent variables. The latent variables processed by VBSPCA can reduce the influence of noise. Dynamic analysis of these latent variables can improve the reliability of fault detection and diagnosis. The effectiveness of the proposed method is verified by an industrial hydrogen production process, and the test results show that both Gaussian prior and Laplace prior-based VBSPCA can effectively detect and diagnose critical faults in AWEs.

The main contributions of this work are summarized as follows:

- 1) This study develops VBSPCA techniques for AWE process diagnosis and fault detection, based on Gaussian prior and Laplace prior, respectively.
- 2) The developed method constructs a sparse prior that naturally reduces the impact of noise and quantifies uncertainty through variational inference.
- 3)  $\ell_1$  regularized vector autoregression extracts dynamic features from latent variables, which can exploit structural information and reduce noise disturbances.

The remainder of the paper is organized as follows. The related works are described in Section 2. The variational Bayesian Sparse PCA method is described in Section 3. The variational inference for these two VBSPCA techniques is presented in Section 4, while dynamic fault detection and diagnosis approach are described in Section 5. Simulation results are provided in Section 6, followed by concluding remarks in Section 7.

## 2. Related work

### 2.1. PCA and Probabilistic PCA

Let  $\mathbf{x}$  denote a sample vector of  $m$  sensors. Assuming that there are  $n$  samples for each sensor, a data matrix  $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{m \times n}$  is composed with each column representing a sample  $\mathbf{x}_i \in \mathbb{R}^m$ . With the projection matrix  $\mathbf{P} \in \mathbb{R}^{m \times r}$ , ( $r < m$ ) satisfies  $\mathbf{P}^\top \mathbf{P} = \mathbf{I}_r$ , the PCA model can be obtained

$$\mathbf{x}_i = \mathbf{P}\mathbf{t}_i + \bar{\mathbf{x}} + \boldsymbol{\xi}_i, \quad i = 1, 2, \dots, n \quad (1)$$

where  $\mathbf{t}_i$  denotes the latent variable with  $r$  dimensions,  $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \in \mathbb{R}^m$  is the mean vector,  $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$  represents additive process noise.

The PCA model can be reformulated by creating matrices  $\mathbf{T} = (\mathbf{t}_1, \mathbf{t}_2, \dots, \mathbf{t}_n) \in \mathbb{R}^{r \times n}$ ,  $\bar{\mathbf{X}} = (\bar{\mathbf{x}}_1, \bar{\mathbf{x}}_2, \dots, \bar{\mathbf{x}}_n) \in \mathbb{R}^{m \times n}$ , and  $\Xi = (\boldsymbol{\xi}_1, \boldsymbol{\xi}_2, \dots, \boldsymbol{\xi}_n) \in \mathbb{R}^{m \times n}$ , PCA can be rewritten as the following

$$\mathbf{X} = \mathbf{P}\mathbf{T} + \bar{\mathbf{X}} + \Xi. \quad (2)$$

In order to maximize the variance of the projected data, PCA can obtain the optimal  $\mathbf{P}$  and  $\mathbf{T}$  by optimizing the following objective function

$$\underbrace{\arg \min}_{\mathbf{P}, \mathbf{T}} \|\mathbf{X} - \bar{\mathbf{X}} - \mathbf{P}\mathbf{T}\|_F^2 \quad \text{s. t. } \mathbf{P}^\top \mathbf{P} = \mathbf{I}_r \quad (3)$$

where  $\|\cdot\|_F$  is the Frobenious norm.

**Remark 1.** In general, the optimal  $\mathbf{P}$  can be obtained by stacking the  $r$  singular vectors corresponding to the first  $r$  largest singular values of the sample covariance matrix  $\mathbf{S} = \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^\top / (n - 1)$ . As a result, latent variable  $\mathbf{t}_i = (\mathbf{x}_i - \bar{\mathbf{x}})\mathbf{P}$  can be obtained[28].

The disadvantage of PCA is that it is not possible to obtain probability distributions with respect to the projection matrix  $\mathbf{P}$  and the latent variable  $\mathbf{t}$ . Probabilistic PCA (PPCA) can be obtained from a Gaussian latent variable model, and its parameters are estimated iteratively using the EM method[29, 30]. The conditional probability distribution of  $\mathbf{x}_i$  can be expressed as

$$p(\mathbf{x}_i | \mathbf{t}_i) = \mathcal{N}(\mathbf{x}_i | \mathbf{P}\mathbf{t}_i + \bar{\mathbf{x}}, \sigma^2 \mathbf{I}). \quad (4)$$

The prior distribution of the latent variable  $\mathbf{t}_i$  is assumed to be a standard Gaussian distribution, the marginal distribution of  $\mathbf{x}_i$  can thus be calculated by integrating out the latent variables

$$p(\mathbf{x}_i) = \int p(\mathbf{x}_i|\mathbf{t}_i)p(\mathbf{t}_i)d\mathbf{t}_i = \mathcal{N}(\mathbf{x}_i|\bar{\mathbf{x}}, \mathbf{P}\mathbf{P}^\top + \sigma^2\mathbf{I}). \quad (5)$$

According to the Maximum likelihood estimation (MLE), the log-likelihood function is given by

$$\mathcal{L}(\mathbf{P}, \bar{\mathbf{x}}, \sigma^2\mathbf{I}|\mathbf{X}) = \sum_{i=1}^n \ln p(\mathbf{x}_i) = \sum_{i=1}^n \ln \mathcal{N}(\mathbf{x}_i|\bar{\mathbf{x}}, \mathbf{P}\mathbf{P}^\top + \sigma^2\mathbf{I}). \quad (6)$$

Maximizing  $\mathcal{L}(\mathbf{P}, \bar{\mathbf{x}}, \sigma^2\mathbf{I}|\mathbf{X})$  yields the optimal  $\mathbf{P}$ ,  $\bar{\mathbf{x}}$ , and  $\sigma^2$ .

**Remark 2.** *Finding the projection matrix  $\mathbf{P}$ , where  $r < m$  accomplishes the reduced dimensionality and feature extraction, is the goal of both PCA and PPCA. In this way, the variable  $\mathbf{x}_i$  can be mapped to the low-dimensional space and the latent variable (extracted feature) can be obtained. This is actually the process of feature extraction.*

Typically, the matrix  $\mathbf{P}$  derived from PCA and PPCA is not sparse, leading to poor interpretability. If  $\mathbf{P}$  is sparse, then the latent variable  $\mathbf{t}$  can be expressed as a weighted combination of a small number of original features, thereby improving the interpretability of the model.

## 2.2. Sparsity analysis

To obtain sparse solutions for  $\mathbf{P}$ , Lasso regression ( $\ell_1$  regularization) and Ridge regression ( $\ell_2$  regularization) are commonly used. The optimization objective of Lasso regression is expressed as

$$\mathbf{P}^* = \underbrace{\arg \min}_{\mathbf{P}} \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{P}\mathbf{t}_i\| + \lambda \|\mathbf{P}\|_1. \quad (7)$$

The optimization objective of Ridge regression is expressed as

$$\mathbf{P}^* = \underbrace{\arg \min}_{\mathbf{P}} \frac{1}{n} \sum_{i=1}^n \|\mathbf{x}_i - \mathbf{P}\mathbf{t}_i\|^2 + \lambda \|\mathbf{P}\|_2^2. \quad (8)$$

**Remark 3.** *Using the  $\ell_1$  norm as a regularization term allows for feature selection, resulting in a sparse model where some regression coefficients can be compressed to zero; while using the  $\ell_2$  norm as a regularization term can help prevent overfitting.*

Following the conditional probability distribution of  $\mathbf{x}_i$  in (4), sparsity analysis will be reformulated as maximizing the posterior probability

$$\begin{aligned} \mathbf{P}^* &= \underbrace{\arg \max}_{\mathbf{P}} \ln \left( \prod_{i=1}^n p(\mathbf{x}_i|\mathbf{t}_i, \mathbf{P})p(\mathbf{P}) \right) \\ &= \underbrace{\arg \max}_{\mathbf{P}} \sum_{i=1}^n \left( \ln p(\mathbf{x}_i|\mathbf{t}_i, \mathbf{P}) + \ln \sum_{i=1}^n p(\mathbf{p}_i) \right) \end{aligned} \quad (9)$$

where  $\ln p(\mathbf{x}_i|\mathbf{t}_i, \mathbf{P}) \propto (\mathbf{x}_i - \mathbf{P}\mathbf{t}_i)^2$ .

It can be seen that the target loss part of (7) is solved here by a probability distribution. The significance of  $p(\mathbf{P})$  is the prior assumption about the probability distribution of  $\mathbf{P}$ . The Ridge Regression will be obtained by assuming that the prior distribution of  $\mathbf{P}$  is a zero-mean Gaussian distribution  $P(\mathbf{p}_i) = \frac{\gamma_i}{\sqrt{2\pi}} \exp(-\gamma_i \|\mathbf{p}_i\|^2)$ . Correspondingly, (9) can be calculated as

$$\mathbf{P}^* = \underbrace{\arg \max}_{\mathbf{P}} \sum_{i=1}^n \left( \ln p(\mathbf{x}_i|\mathbf{t}_i, \mathbf{P}) + \gamma \sum_{i=1}^n \|\mathbf{p}_i\|^2 \right). \quad (10)$$

The Lasso Regression will be obtained by assuming that the prior distribution of  $\mathbf{P}$  is a zero-mean Laplace distribution  $P(\mathbf{p}_i) = \frac{\varphi_i}{2} \exp(-\varphi_i |\mathbf{p}_i|)$ , Correspondingly, (9) can be calculated as

$$\mathbf{P}^* = \underbrace{\arg \max}_{\mathbf{P}} \sum_{i=1}^n \left( \ln p(\mathbf{x}_i|\mathbf{t}_i, \mathbf{P}) + \varphi \sum_{i=1}^n |\mathbf{p}_i| \right). \quad (11)$$

Consequently, it can be shown that in Bayesian inference, the Laplace prior is in fact equivalent to  $\ell_1$  regularization, and the Gaussian prior is equivalent to  $\ell_2$  regularization. It can usually be solved by EM methods, however, as a point estimation method, the EM method can only estimate the maximum posterior of the single most probable value for each parameter. This parameter learning method may lead to local optima and overfitting, thus limiting the performance of the model.

### 3. Variational Bayesian Sparse PCA

#### 3.1. Gaussian prior-based VBSPCA

Suppose  $\mathbf{X}$  can be decomposed into a low-rank matrix  $\mathbf{H}$  and the noise  $\Xi$

$$\mathbf{X} = \mathbf{H} + \boldsymbol{\xi} = \sum_{i=1}^k \mathbf{p}_i \mathbf{t}_i^\top + \Xi \quad (12)$$

where  $\mathbf{H} = \mathbf{P}\mathbf{T}^\top$  such that  $\text{Rank}(\mathbf{H}) = r \ll \min(n, m)$ .

According to the principle of  $\ell_2$  regularized sparsity, a sparsity inducing Gaussian priors is used on each coefficient of the unknown vector based on Automatic relevance determination (ARD). The columns of  $\mathbf{P}$  and  $\mathbf{T}$  are associated with Gaussian priors of precisions  $\gamma$  to ensure that the majority of the columns in each are set to zero, so  $\gamma$  is closely related to the sparsity of the solution[21]

$$\begin{aligned} p(\mathbf{P}|\gamma) &= \prod_{i=1}^k \mathcal{N}(\mathbf{p}_i|0, \gamma_i^{-1} \mathbf{I}_m), \\ p(\mathbf{T}|\gamma) &= \prod_{i=1}^k \mathcal{N}(\mathbf{t}_i|0, \gamma_i^{-1} \mathbf{I}_n) \end{aligned} \quad (13)$$

where  $\mathbf{I}_m$  is the  $m \times m$  identity matrix,  $\mathbf{I}_n$  is the  $n \times n$  identity matrix. Thus, the columns of  $\mathbf{P}$  and  $\mathbf{T}$  have the same sparsity profile enforced by the common precisions  $\gamma_i$ .

In this inference process, many of the precisions  $\gamma_i$  will assume very large values, which effectively removes the corresponding outer-products from  $\mathbf{H}$  and thus reduces the rank of the estimate. Gamma priors  $p(\gamma_i)$  can be placed on the precision parameters of the above Gaussian distributions of  $\mathbf{P}$  and  $\mathbf{T}$ . The form of conjugate Gamma hyperprior is expressed as

$$p(\gamma_i) = \text{Gam}(a_0, \frac{1}{b_0}) \propto \gamma_i^{a_0-1} \exp(-b_0 \gamma_i) \quad (14)$$

**Remark 4.** *To ensure that the hyperpriors are sufficiently vague and diffuse, we treat the parameters  $a_0$  and  $b_0$  as fixed constants and assign them very small values, such as  $10^{-5}$ , instead of estimating them from the data.*

**Remark 5.** *The noise in the measurements is assumed to be i.i.d. drawn from a Gaussian distribution, which has an effect on each of the measurements. Noting that this construction is only within the prior, and that an approximation to the posterior  $\boldsymbol{\xi}$  can be inferred for perturbations to the prior.*

Mathematically, the noise in (2) is modeled as

$$p(\boldsymbol{\Xi}|\beta) = \prod_{j=1}^m \prod_{i=1}^n \mathcal{N}(\boldsymbol{\Xi}_{ji}|0, \beta^{-1}). \quad (15)$$

The parameter selection of the prior  $\alpha$  can largely affect the parameter estimation of the posterior distribution. To reduce the effect of the parameter selection of the original distribution, the noise precision  $\beta$  is set to the non-informative Jeffrey's prior

$$p(\beta) = \beta^{-1}. \quad (16)$$

In accordance with the form of PCA in (1), the mean  $\bar{\mathbf{X}}$  in (12) is modeled as

$$p(\bar{\mathbf{X}}|\alpha) = \prod_{j=1}^m \prod_{i=1}^n \mathcal{N}(\bar{\mathbf{X}}_{ji}|0, \alpha_{ji}^{-1}). \quad (17)$$

**Remark 6.** *In classical PCA fault detection method, a preliminary step includes zero-centering. Hence, the decomposition of  $\mathbf{X}$  doesn't consider the mean values. For the rigor of the algorithm, we give here the form of the mean  $\bar{\mathbf{X}}$ , which can also be interpreted as the sparse error.*

Similarly, the precision  $\alpha$  is set to the non-informative Jeffrey's prior

$$p(\alpha_{ji}) = \alpha_{ji}^{-1}. \quad (18)$$

The conditional probability distribution for  $\mathbf{X}$  can be expressed as follows

$$p(\mathbf{X}|\mathbf{P}, \mathbf{T}, \beta) = \prod_{j=1}^m \prod_{i=1}^n \mathcal{N}(y_{ji}|x_{ji}, \beta^{-1}). \quad (19)$$

Finally, with the conditional probability and priors, the joint distribution is expressed as

$$p(\mathbf{X}, \mathbf{P}, \mathbf{T}, \bar{\mathbf{X}}, \boldsymbol{\gamma}, \beta) = p(\mathbf{X}|\mathbf{P}, \mathbf{T}, \bar{\mathbf{X}}, \beta) p(\mathbf{P}|\boldsymbol{\gamma}) p(\mathbf{T}|\boldsymbol{\gamma}) p(\bar{\mathbf{X}}|\boldsymbol{\alpha}) p(\boldsymbol{\alpha}) p(\boldsymbol{\gamma}) p(\beta). \quad (20)$$

### 3.2. Laplace prior-based VBSPCA

The Laplace distribution has coarser tails than the Gaussian distribution, hence the name heavy-tailed distribution. It means that the Laplace distribution can fit heavy noise and outliers better than the Gaussian distribution.

**Remark 7.** *It is necessary to employ another prior, a two-level hierarchical decomposition of the Laplace distribution, as the Laplace distribution is not a conjugate prior to the Gaussian distribution and does not support posterior inference within a Bayesian framework.*

According to the principle of  $\ell_2$  regularized sparsity, the first level of  $\mathbf{P}$  introduces a Gaussian prior  $p(\mathbf{P}_{ji}|\boldsymbol{\eta}_{ji}) \sim \mathcal{N}(0, \boldsymbol{\eta}_{ji})$ , while the second level of  $\mathbf{P}$  is an exponential distribution hyperprior to the variance  $\boldsymbol{\eta}_{ji}$ ,  $p(\boldsymbol{\eta}_{ji}) = \frac{1}{\varphi} \exp(-\frac{\boldsymbol{\eta}_{ji}}{\varphi})$

$$\begin{aligned} p(\mathbf{P}_{ji}|\boldsymbol{\eta}_{ji}) &= \int p(\mathbf{P}_{ji}|\boldsymbol{\eta}_{ji}) p(\boldsymbol{\eta}_{ji}) d\boldsymbol{\eta}_{ji} \\ &= \frac{1}{2} \sqrt{\frac{2}{\varphi}} \exp(-\sqrt{\frac{2}{\varphi}} |\mathbf{P}_{ji}|) \end{aligned} \quad (21)$$

Assuming that the noise  $\boldsymbol{\xi}$  follows a Gaussian distribution  $\boldsymbol{\xi} \sim \mathcal{N}(0, \boldsymbol{\vartheta}^{-1})$ , a prior over the inverse noise variance  $\boldsymbol{\vartheta}$  is introduced

$$p(\boldsymbol{\vartheta}) = \text{Gam}(\boldsymbol{\vartheta}|c_0, d_0). \quad (22)$$

In accordance with the form of PCA in (1), it is assumed that both the mean  $\bar{\mathbf{x}}$  and the latent variable  $\mathbf{t}$  have Gaussian distributions

$$\begin{aligned} p(\bar{\mathbf{x}}) &\sim \mathcal{N}(0, \varsigma^{-1} \mathbf{I}) \\ p(\mathbf{t}_i) &\sim \mathcal{N}(0, \mathbf{I}) \end{aligned} \quad (23)$$

Finally, with the conditional probability and priors, the joint distribution is expressed as

$$p(\mathbf{X}, \mathbf{P}, \mathbf{t}, \bar{\mathbf{x}}, \boldsymbol{\vartheta}) = \prod_{i=1}^n p(\mathbf{y}|\mathbf{P}, \mathbf{t}, \bar{\mathbf{x}}, \boldsymbol{\vartheta}, \boldsymbol{\eta}) P(\mathbf{t}) p(\mathbf{P}|\boldsymbol{\eta}) p(\boldsymbol{\eta}) p(\bar{\mathbf{x}}) p(\boldsymbol{\vartheta}) \quad (24)$$



## 4. Variational Bayesian inference

### 4.1. A review of variational inference

A variational Bayesian approach is implemented in this section to approximate the posterior distribution of the parameters. The variational Bayesian of learning  $p(\boldsymbol{\Theta}|\mathbf{y})$  avoids intractable integration by learning an approximating distribution  $Q(\boldsymbol{\Theta})$ . As a result, the marginal log-likelihood function  $\ln p(\mathbf{y})$  can be separated into two components: the Evidence Lower Bound (ELBO)  $\mathcal{L}(Q(\boldsymbol{\Theta}))$  and Kullback-Leibler (KL) divergence  $\mathbb{KL}(Q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\mathbf{y}))$ [27]

$$\begin{aligned}\ln p(\mathbf{y}) &= \mathcal{L}(Q(\boldsymbol{\Theta})) + \mathbb{KL}(Q(\boldsymbol{\Theta})||p(\boldsymbol{\Theta}|\mathbf{y})) \\ &= \ln \int_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}) \left\{ \frac{p(\boldsymbol{\Theta}, \mathbf{y})}{Q(\boldsymbol{\Theta})} \right\} d\boldsymbol{\Theta} - \int_{\boldsymbol{\Theta}} Q(\boldsymbol{\Theta}) \ln \left\{ \frac{p(\boldsymbol{\Theta}, \mathbf{y})}{Q(\boldsymbol{\Theta})} \right\} d\boldsymbol{\Theta}\end{aligned}\quad (25)$$

where  $Q(\boldsymbol{\Theta})$  approximates as closely as possible the true distribution  $p(\boldsymbol{\Theta}|\mathbf{y})$  through KL divergence.

**Remark 8.** Since  $KL(Q \parallel p) \geq 0$ ,  $\mathcal{L}(Q(\boldsymbol{\Theta}))$  is a rigorous lower bound for  $\ln p(\mathbf{y})$ . Maximizing  $\mathcal{L}(Q(\boldsymbol{\Theta}))$  is equivalent to minimizing  $KL(Q \parallel p)$ , and thus the posterior distribution  $p(\boldsymbol{\Theta}|\mathbf{y})$  can be approximated by  $Q(\boldsymbol{\Theta})$ .

With the Mean field variational inference (MFVI), the free joint distribution is chosen to be factorable [31]

$$Q(\boldsymbol{\Theta}) = \prod_{i=1}^I q_i(\Theta_i). \quad (26)$$

In order to find an approximate posterior distribution for each latent variable, the expected values of all parameters in the joint distribution are relative to their most recent distribution. The posterior approximation  $q_i(\Theta_i)$  for each latent variable  $\Theta_i$  is optimized by

$$q_j(\Theta_j) = \frac{\exp \mathbb{E}_{i \neq j}(\ln p(\mathbf{y}, \boldsymbol{\Theta}))}{\int_{\Theta_j} \exp(\mathbb{E}_{i \neq j}(\ln p(\mathbf{y}, \boldsymbol{\Theta})))}. \quad (27)$$

where  $\mathbb{E}(\cdot)$  is the expectation.

### 4.2. Variational inference for Gaussian prior-based VBSPCA

The latent variable and parameters of the Gaussian prior-based VBSPCA are  $\boldsymbol{\Theta} = (\mathbf{P}, \mathbf{T}, \bar{\mathbf{X}}, \boldsymbol{\gamma}, \boldsymbol{\alpha}, \beta)$ . Furthermore, following (26),  $\boldsymbol{\Theta}$  can be written as

$$Q(\boldsymbol{\Theta}) = q(\mathbf{P})q(\mathbf{T})q(\bar{\mathbf{X}})q(\boldsymbol{\gamma})q(\boldsymbol{\alpha})q(\beta). \quad (28)$$

Optimizing the approximate distribution  $Q(\Theta)$  based on the joint distribution of (20) gives the form of these parameters and latent variables

$$q(\mathbf{p}_i) = \mathcal{N}(\mathbf{p}_i, |\boldsymbol{\mu}_i^p, \boldsymbol{\Sigma}_i^p) \quad (29a)$$

$$q(\mathbf{t}_j) = \mathcal{N}(\mathbf{t}_j, |\boldsymbol{\mu}_j^t, \boldsymbol{\Sigma}_j^t) \quad (29b)$$

$$q(\bar{\mathbf{X}}_{ji}) = \mathcal{N}(\bar{\mathbf{X}}_{ji}, |\boldsymbol{\mu}_{ji}^{\bar{\mathbf{X}}}, \boldsymbol{\Sigma}_{ji}^{\bar{\mathbf{X}}}) \quad (29c)$$

$$q(\gamma) = \text{Gam}(\gamma_i | a, b) \quad (29d)$$

Following (16) and (18),  $\alpha$  and  $\beta$  introduce a non-informative Jeffrey's prior with the following expectations

$$\begin{aligned} \mathbb{E}(\alpha_{ji}) &= \frac{\mathbb{E}(\beta) + \mathbb{E}(\alpha_{ji})}{1 + \mathbb{E}(\bar{\mathbf{X}}_{ji})(\mathbb{E}(\beta) + \mathbb{E}(\alpha_{ji}))} \\ \mathbb{E}(\beta) &= \frac{mn}{\mathbb{E}(\|\mathbf{X} - \mathbf{P}\mathbf{T}^\top - \bar{\mathbf{X}}\|_F^2)} \end{aligned} \quad (30)$$

The approximate distribution  $q(\mathbf{p}_i)$  in (29a) is given by

$$\begin{aligned} \boldsymbol{\mu}_i^p &= \mathbb{E}(\beta) \boldsymbol{\Sigma}_i^p \mathbb{E}(\mathbf{T})(\mathbf{y}_i - \mathbb{E}(\bar{\mathbf{x}}_i)) \\ \boldsymbol{\Sigma}_i^p &= (\mathbb{E}(\beta) \mathbb{E}(\mathbf{T}^\top \mathbf{T}) + \mathbb{E}(\gamma) \mathbf{I})^{-1}. \end{aligned} \quad (31)$$

The approximate distribution  $q(\mathbf{t}_i)$  in (29b) is given by

$$\begin{aligned} \boldsymbol{\mu}_j^t &= \mathbb{E}(\beta) \boldsymbol{\Sigma}_j^t \mathbb{E}(\mathbf{P})(\mathbf{y}_j - \mathbb{E}(\bar{\mathbf{x}}_j)) \\ \boldsymbol{\Sigma}_j^t &= (\mathbb{E}(\beta) \mathbb{E}(\mathbf{P}^\top \mathbf{P}) + \mathbb{E}(\gamma) \mathbf{I})^{-1}. \end{aligned} \quad (32)$$

The approximate distribution  $q(\bar{\mathbf{X}}_{ji})$  in (29c) is given by

$$\begin{aligned} \boldsymbol{\mu}_{ji}^{\bar{\mathbf{X}}} &= \frac{\mathbb{E}(\beta)}{\mathbb{E}(\beta) + \mathbb{E}(\alpha_{ji})} (\mathbf{X}_{ji} - \mathbb{E}(\mathbf{p}_i) \mathbb{E}(\mathbf{t}_i)^\top) \\ \boldsymbol{\Sigma}_{ji}^{\bar{\mathbf{X}}} &= \frac{1}{\mathbb{E}(\beta) + \mathbb{E}(\alpha_{ji})}. \end{aligned} \quad (33)$$

The approximate distribution  $q(\gamma)$  in (29d) is given by

$$\begin{aligned} a &= a_0 + \frac{m+n}{2} \\ b &= b_0 + \frac{\mathbb{E}(\mathbf{p}_i^\top \mathbf{p}_i) + \mathbb{E}(\mathbf{t}_i^\top \mathbf{t}_i)}{2}. \end{aligned} \quad (34)$$

The required moments are easily evaluated using basic properties of probability distributions

$$\begin{aligned} \mathbb{E}(\mathbf{P}^\top \mathbf{P}) &= \mathbb{E}(\mathbf{P})^\top \mathbb{E}(\mathbf{P}) + m \boldsymbol{\Sigma}_i^p, \\ \mathbb{E}(\mathbf{T}^\top \mathbf{T}) &= \mathbb{E}(\mathbf{T})^\top \mathbb{E}(\mathbf{T}) + n \boldsymbol{\Sigma}_i^t, \\ \mathbb{E}(\mathbf{p}_i^\top \mathbf{p}_i) &= \mathbb{E}(\mathbf{p}_i)^\top \mathbb{E}(\mathbf{p}_i) + m (\boldsymbol{\Sigma}_i^p)_{ii}, \\ \mathbb{E}(\mathbf{t}_i^\top \mathbf{t}_i) &= \mathbb{E}(\mathbf{t}_i)^\top \mathbb{E}(\mathbf{t}_i) + n (\boldsymbol{\Sigma}_i^t)_{ii}, \end{aligned} \quad (35)$$

$$\begin{aligned} \mathbb{E}(\|\mathbf{X} - \mathbf{P}\mathbf{T}^\top - \bar{\mathbf{X}}\|_F^2) &= \|\mathbf{X} - \mathbb{E}(\mathbf{P}) \mathbb{E}(\mathbf{T})^\top - \mathbb{E}(\bar{\mathbf{X}}_{ji})\|_F^2 + \text{tr}(n \mathbb{E}(\mathbf{P})^\top \mathbb{E}(\mathbf{P}) \boldsymbol{\Sigma}_i^t) \\ &\quad + \text{tr}(m \mathbb{E}(\mathbf{T})^\top \mathbb{E}(\mathbf{T}) \boldsymbol{\Sigma}_i^p) + \text{tr}(mn \boldsymbol{\Sigma}_i^p \boldsymbol{\Sigma}_i^t) + \sum_{i=1}^n \sum_{j=1}^m \boldsymbol{\Sigma}_{ji}^{\bar{\mathbf{X}}}. \end{aligned}$$

#### 4.3. Variational inference for VBSPCA with Laplace prior

Following (24), the latent variables and parameters of the Laplace distribution-based VBSPCA are  $\Theta = (\mathbf{P}, \mathbf{t}, \bar{\mathbf{x}}, \boldsymbol{\vartheta}, \boldsymbol{\eta})$ . Furthermore,  $\Theta$  can be written as

$$Q(\Theta) = q(\mathbf{P})q(\mathbf{t})q(\bar{\mathbf{x}})q(\boldsymbol{\vartheta})q(\boldsymbol{\eta}). \quad (36)$$

Optimizing the approximate distribution  $Q(\Theta)$  based on the joint distribution of (24) gives the form of these parameters and latent variables

$$q(\mathbf{P}) = \prod_{i=1}^r \mathcal{N}(\mathbf{P}_i, |\boldsymbol{\mu}_i^{\mathbf{P}}, \boldsymbol{\Sigma}_i^{\mathbf{P}}|) \quad (37a)$$

$$q(\mathbf{t}) = \prod_{i=1}^n \mathcal{N}(\mathbf{t}_i, |\boldsymbol{\mu}_i^{\mathbf{t}}, \boldsymbol{\Sigma}_i^{\mathbf{t}}|) \quad (37b)$$

$$q(\bar{\mathbf{x}}) = \mathcal{N}(\bar{\mathbf{x}}, |\boldsymbol{\mu}^{\bar{\mathbf{x}}}, \boldsymbol{\Sigma}^{\bar{\mathbf{x}}}|) \quad (37c)$$

$$q(\boldsymbol{\vartheta}) = \text{Gam}(\boldsymbol{\vartheta}|c, d) \quad (37d)$$

$$q(\boldsymbol{\eta}) = \frac{1}{\sqrt{\pi\boldsymbol{\eta}_{ij}\varphi}} \exp\left(-\frac{1}{2\boldsymbol{\eta}_{ji}}(\mathbf{P}_{ji})^\top(\mathbf{P}_{ji}) - \frac{\boldsymbol{\eta}_{ji}}{\varphi} + \sqrt{\frac{2}{\varphi}}|\mathbf{P}_{ji}|\right) \quad (37e)$$

The approximate distribution  $q(\mathbf{P})$  in (37a) is given by

$$\begin{aligned} \boldsymbol{\mu}_i^{\mathbf{P}} &= \mathbb{E}(\boldsymbol{\vartheta})\boldsymbol{\Sigma}_i^{\mathbf{P}} \sum_{i=1}^n \mathbb{E}(\mathbf{x}_i)(\mathbf{y}_i - \mathbb{E}(\bar{\mathbf{x}})) \\ \boldsymbol{\Sigma}_i^{\mathbf{P}} &= \left( \mathbb{E}(\boldsymbol{\vartheta}) \sum_{i=1}^n \mathbb{E}(\mathbf{x}_i \mathbf{x}_i^\top) \text{dg}(\mathbb{E}(\boldsymbol{\eta}_i)) + \mathbf{I} \right)^{-1} \text{dg}(\mathbb{E}(\boldsymbol{\eta}_i)). \end{aligned} \quad (38)$$

where dg converts the vector to a diagonal matrix.

The approximate distribution  $q(\mathbf{t})$  in (37b) is given by

$$\begin{aligned} \boldsymbol{\mu}_i^{\mathbf{t}} &= \mathbb{E}(\boldsymbol{\vartheta})\boldsymbol{\Sigma}_i^{\mathbf{t}}\mathbb{E}(\mathbf{P})(\mathbf{y}_i - \mathbb{E}(\bar{\mathbf{x}})) \\ \boldsymbol{\Sigma}_i^{\mathbf{t}} &= (\mathbb{E}(\boldsymbol{\vartheta})\mathbb{E}(\mathbf{P}^\top \mathbf{P}) + \mathbf{I})^{-1}. \end{aligned} \quad (39)$$

The approximate distribution  $q(\bar{\mathbf{x}})$  in (37c) is given by

$$\begin{aligned} \boldsymbol{\mu}^{\bar{\mathbf{x}}} &= \mathbb{E}(\boldsymbol{\vartheta})\boldsymbol{\Sigma}^{\bar{\mathbf{x}}} \sum_{i=1}^n (\mathbf{y}_i - \mathbb{E}(\mathbf{P})\mathbb{E}(\mathbf{x}_i)) \\ \boldsymbol{\Sigma}^{\bar{\mathbf{x}}} &= (n\mathbb{E}(\boldsymbol{\vartheta}) + \varsigma)^{-1} \mathbf{I}. \end{aligned} \quad (40)$$

The approximate distribution  $q(\boldsymbol{\vartheta})$  in (37d) is given by

$$\begin{aligned} c &= c_0 + \frac{nm}{2} \\ d &= d_0 + \frac{1}{2} \sum_{i=1}^n \|\bar{\mathbf{x}}\|^2 + \text{tr}(\mathbb{E}(\mathbf{P}^\top \mathbf{P})\mathbb{E}(\mathbf{x}_i \mathbf{x}_i^\top)) \\ &\quad + 2\mathbb{E}(\bar{\mathbf{x}})^\top \mathbb{E}(\mathbf{P})\mathbb{E}(\mathbf{x}_i) - 2\mathbf{y}_i^\top \mathbb{E}(\mathbf{P})\mathbb{E}(\mathbf{x}_i) - 2\mathbf{y}_i^\top \mathbb{E}(\bar{\mathbf{x}}) + \|\mathbf{y}_i\|^2. \end{aligned} \quad (41)$$

The approximate distribution  $q(\boldsymbol{\eta})$  (37e) is given by

$$\mu_{\boldsymbol{\eta}_{ji}} = \frac{1}{2}(\varphi + \sqrt{2\varphi}|\mathbf{P}_{ji}|). \quad (42)$$

## 5. Dynamic Process Monitoring

### 5.1. Fault detection

In this section, fault detection is accomplished and the dynamics of the latent variables are examined after the projection loading matrix  $\mathbf{P}$  is calculated by variational inference using the VBSPCA approach, which is based on Gaussian prior and Laplace prior.

**Remark 9.** *Fault detection and diagnosis depends more on the dynamics of the latent variables  $\mathbf{t}_k^*$  of the newly collected samples, although it also depends on the projection matrix  $\mathbf{P}$ . Therefore, it is more interpretable to analyze the dynamics of the latent variables directly.*

A new collected sample  $\mathbf{y}_k \in \mathbb{R}^m$  is obtained at time  $k$  with time interval  $L$ , the latent variable  $\mathbf{t}_k^*$  of the new collected sample  $\mathbf{y}_k$  is obtained by a linear transformation of  $\mathbf{P}$

$$\mathbf{t}_k^* = \mathbf{P}^\top \mathbf{y}_k \quad (43)$$

where  $\mathbf{t}_k^*$  is a potential time series, which means that  $\mathbf{t}_k^*$  at the current moment is influenced by the previous moment.

As the dynamic relationship of the latent variable  $\mathbf{t}_k^*$  is not explained, it is necessary to build a dynamic model to characterize the auto correlation. As a result,  $\mathbf{t}_k^*$  can be expressed as the following VAR model

$$\mathbf{t}_k^* = \sum_{d=1}^{\tau} \boldsymbol{\omega}_d \mathbf{t}_{k-d}^* + \boldsymbol{\epsilon}_k, k = \tau + 1, \dots, L \quad (44)$$

where  $\tau$  is the order of dynamics,  $\boldsymbol{\epsilon}_k$  is zero mean Gaussian noise,  $\boldsymbol{\omega}_i \in \mathbb{R}^{r \times r}, d = 1, 2, \dots, \tau$  represents the coefficient matrix of the VAR model.

The objective is to learn the coefficient matrix  $\boldsymbol{\omega}$  from the latent variable, and the new linear expression for  $\boldsymbol{\omega}$  and  $\boldsymbol{\chi}_k$  is rewritten as

$$\begin{aligned} \boldsymbol{\omega} &= [\boldsymbol{\omega}_1, \dots, \boldsymbol{\omega}_\tau]^\top \in \mathbb{R}^{(r \cdot \tau) \times r} \\ \boldsymbol{\chi}_k &= [\mathbf{t}_{k-1}^{*\top}, \dots, \mathbf{t}_{k-\tau}^{*\top}]^\top \in \mathbb{R}^{(r \cdot \tau)} \end{aligned} \quad (45)$$

The VAR model for the latent variable  $\mathbf{t}$  can be rewritten as

$$\mathbf{t}_k^* \approx \sum_{d=1}^{\tau} \boldsymbol{\omega}_d \mathbf{t}_{k-d}^* = \boldsymbol{\omega}^\top \boldsymbol{\chi}_k, k = \tau + 1, \dots, L \quad (46)$$

In high dimensional scaling, where the transition matrix driving the temporal dynamics displays a more complex structure. In order to train the sparse projection in the regression model,  $\ell_1$  regularization is therefore required

$$\hat{\boldsymbol{\omega}} = \underbrace{\arg \min}_{\boldsymbol{\omega}} \frac{1}{2} \|\mathbf{Z} - \mathbf{Q}\boldsymbol{\omega}\|_2^2 + \lambda \|\boldsymbol{\omega}\|_1 \quad (47)$$

where  $\mathbf{Z}$  and  $\mathbf{Q}$  are defined as  $\mathbf{Z} = [\mathbf{t}_{\tau+1}^\top, \dots, \mathbf{t}_L^\top]^\top \in \mathbb{R}^{(F-\tau) \times r}$ ,  $\mathbf{Q} = [\mathbf{x}_{\tau+1}^\top, \dots, \mathbf{x}_L^\top]^\top \in \mathbb{R}^{(F-\tau) \times (r \cdot \tau)}$ , respectively.

**Remark 10.** A sparse VAR about  $\boldsymbol{\omega}$  is built such that  $\boldsymbol{\omega}$  becomes a sparse matrix for feature selection on the latent variable  $\mathbf{t}_k^*$  of the newly collected sample  $\mathbf{y}_k$ . This further reduces the noise perturbation of the newly collected sample  $\mathbf{y}_k$  and improves the model interpretability.

The prediction of  $\hat{\mathbf{t}}_k$  from the sparse VAR model is

$$\hat{\mathbf{t}}_k = \sum_{d=1}^{\tau} \hat{\boldsymbol{\omega}}_d \mathbf{t}_{k-d}^* = \sum_{d=1}^{\tau} \hat{\boldsymbol{\omega}}_d \mathbf{P}^\top \mathbf{y}_{k-d}, k = \tau + 1, \dots, L. \quad (48)$$

From the predicted latent variable  $\hat{\mathbf{t}}_k$ , the  $T_2$  statistics and  $SPE$  statistics can be obtained, which are the essential indicators for performing fault detection.

The  $T_2$  statistic can be constructed as

$$T_2 = \hat{\mathbf{t}}_k \boldsymbol{\Lambda} \hat{\mathbf{t}}_k^\top. \quad (49)$$

where  $\boldsymbol{\Lambda}$  is a diagonal matrix.

The  $SPE$  statistic can be constructed as

$$SPE = \|(\mathbf{I} - \mathbf{P}\mathbf{P}^\top) \hat{\mathbf{t}}_k\|^2. \quad (50)$$

In this work, the kernel density estimation (KDE) method is used to calculate the control limit, which is a non-parametric method to build an unknown density function

$$\hat{f}_h(e) = \frac{1}{L} \sum_{i=1}^L K_h(e - e_i) = \frac{1}{Lh} \sum_{i=1}^L K_h\left(\frac{e - e_i}{h}\right). \quad (51)$$

where  $K_h$  is a non-negative kernel function and  $h$  is a relative smoothing parameter called bandwidth.  $e$  is the reconstruction error, to construct the density estimate accordingly. The confidence limit can be obtained with a confidence level  $\alpha = 0.95$ . When the statistic exceeds the confidence limit, the test sample  $\mathbf{y}_k$  is considered faulty.

## 5.2. Fault diagnosis

A detectable fault should have an impact on the measurement vector that deviates from the normal scenario. The measurement vector for the fault-free part is denoted by  $\boldsymbol{\psi}$  and will be unavailable when a fault occurs[32]. In the presence of a process fault,  $\mathbf{y}$  is represented as

$$\mathbf{y} = \boldsymbol{\psi} + \mathbf{f}\boldsymbol{\phi} \quad (52)$$

where  $\phi$  is orthonormal so that  $\|\mathbf{f}\|$  represents the magnitude of the fault.  $\psi$  is the fault-free part of the measurement.

The objective of fault reconstruction is to estimate the fault-free part  $\psi$  by removing the effects of faults. A reconstructed sample vector  $\psi_k \in \psi$  is calculated by correcting the effect of a fault on the process data  $\mathbf{y}$

$$\psi_k = \mathbf{y} - f_k \phi_k \quad (53)$$

The fault detection index for reconstructing variables can be written as

$$\text{Index}(\psi_k) = \mathbf{y}_k^\top \Phi \mathbf{y}_k = \|\mathbf{y}_k\|_\Phi^2 = \|\psi - f_k \phi_k\|_\Phi^2. \quad (54)$$

The purpose of reconstruction is to find the value that minimizes  $f_k$ . Taking the derivation of (54),  $f_k = (\phi_k^\top \Phi \phi_k)^{-1} \phi_k^\top \Phi \mathbf{y}$  can be obtained. The reconstruction-based contribution  $\text{RBC}_k^{\text{Index}}$  along the fault direction can be expressed as

$$\text{RBC}_k^{\text{Index}} = \|\phi_k f_k\|_\mathbf{y}^2 = \psi^\top \Phi \phi_k (\phi_k^\top \Phi \phi_k)^{-1} \phi_k^\top \Phi \mathbf{y}. \quad (55)$$

The reconstructed index  $\text{Index}(\psi_k)$  in (54) can be expressed as

$$\text{Index}(\psi_k) = \mathbf{y}^\top \Phi \left[ \mathbf{I} - \phi_k (\phi_k^\top \Phi \phi_k)^{-1} \phi_k^\top \Phi \right] \mathbf{y} = \mathbf{y}^\top \Phi \mathbf{y} - \text{RBC}_k^{\text{Index}} \quad (56)$$

As a result,  $\text{Index}(\mathbf{x}) = \text{Index}(\psi_k) + \text{RBC}_k^{\text{Index}}$ .

For  $T_2$  indexes,  $\text{RBC}_k^{T_2}$  can be expressed as

$$\text{RBC}_k^{T_2} = \frac{(\phi_k^\top \mathbf{P} \Lambda^{-1} \mathbf{P}^\top \mathbf{y})^2}{(\mathbf{P} \Lambda^{-1} \mathbf{P}^\top)_{kk}} \quad (57)$$

where  $(\mathbf{P} \Lambda^{-1} \mathbf{P}^\top)_{kk}$  is the  $i$ th diagonal element of  $(\mathbf{P} \Lambda^{-1} \mathbf{P}^\top)$ .

For  $SPE$  indexes,  $\text{RBC}_k^{SPE}$  can be expressed as

$$\text{RBC}_k^{SPE} = \frac{(\phi_k^\top (\mathbf{I} - \mathbf{P}^\top \mathbf{P}) \mathbf{y})^2}{(\mathbf{I} - \mathbf{P}^\top \mathbf{P})_{kk}} \quad (58)$$

where  $(\mathbf{I} - \mathbf{P}^\top \mathbf{P})_{kk}$  is the  $i$ th diagonal element of  $(\mathbf{I} - \mathbf{P}^\top \mathbf{P})$ .

## 6. Industrial AWE hydrogen production studies

The AWE hydrogen production industrial process consists of three main parts, electrolysis module, gas-liquid separation module, and circulation module. As shown in Figure 1, they include alkaline electrolyzer, heat exchanger, gas-liquid separator, condenser and dryer. The direct current supply powers the alkaline water electrolyzer, and the electrolyte is then decomposed to generate

hydrogen and oxygen. Hydrogen and oxygen are dissolved in the electrolyte. When the pressure rises to the rated value, the mixtures is sent to the gas-liquid separator through the back pressure valve, where gas and electrolyte are separated. After separation, the gases and liquid go through different routes. The filtered electrolyte is cooled to the working temperature by the condenser, and then the electrolyte is returned to the electrolytic cell through the circulating pump, which can reduce the temperature of the electrolyzer. The separated gas is cooled by the condenser, and the cooled gas enters the gas dryer for dehydration. In order to improve the purity of hydrogen, hydrogen is purified and stored by  $H_2$  purification equipment. The liquid flows through the circulation pump and filter and then returns to the electrolysis cell from the bottom of the separators. The hydrogen is tested for purity and then collected and stored. During the electrolysis process, the circulation module cools the system while circulating the electrolyte. Pure water is consumed by electrolysis, so it is necessary to supplement water to get electrolyte. The generated electrolyte and the separated electrolyte are circulated into the electrolyzer to maintain the set concentration.

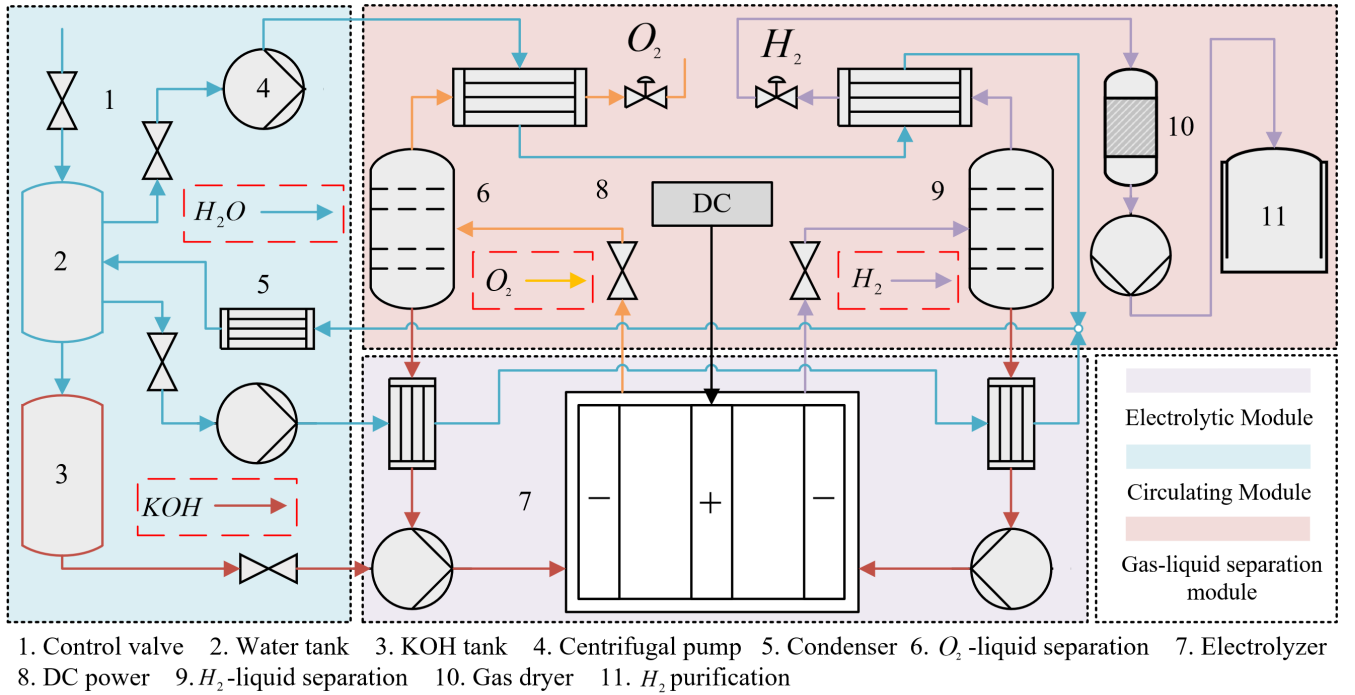


Figure 1: Flow chart of alkaline electrolysis system.

The data of hydrogen production process used in this study comes from the real data of a large-scale hydrogen production plant, including normal data and fault data. According to privacy requirements of the plant, the raw data is desensitized in order to ensure the privacy and security of the information. The variables of the electrolytic water hydrogen production process can be divided into control variables and process variables, and the regulation and monitoring of the variables can be realized by the automated control system to ensure the stability, efficiency and safety of the electrolytic water hydrogen production process. The plant collects 32 variables, including 10 control

variables and 22 process variables. Table 1 lists the control variables and Table 2 lists the process variables. The main faults that may occur in a AWE hydrogen production system are summarized in the Table 3. Ten common faults are present in all three components of the AWE hydrogen production system.

Table 1: Control variable in AWE system.

Control variable	Variable description	Unit
CV(1)	Hydrogen mass flow	$km^3/h$
CV(2)	Oxygen mass flow	$km^3/h$
CV(3)	Hydrogen gas-liquid separator liquid level	L
CV(4)	Oxygen gas-liquid separator liquid level	L
CV(5)	Water tank return flow	$km^3/h$
CV(6)	KOH tank outlet flow	$km^3/h$
CV(7)	Electrolytic cell electrode current	A
CV(8)	Electrolyzer total voltage	V
CV(9)	Electrolyte concentration	wt%
CV(10)	Electrolyzer inlet flow	$km^3/h$



Table 2: Process variable in AWE system.

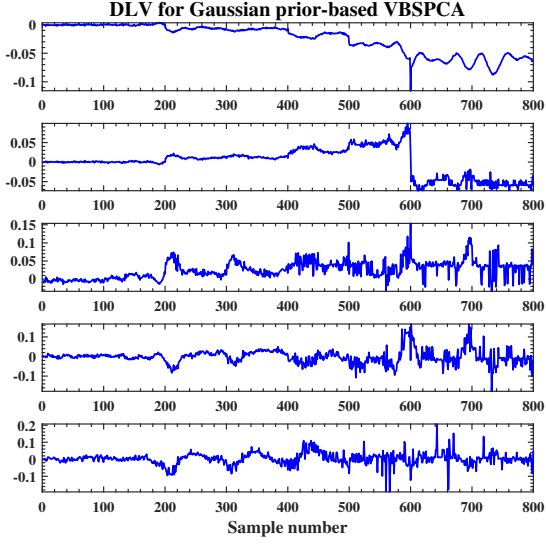
Process variable	Variable description	Unit
PV(1)	Electrolyzer cell temperature	°C
PV(2)	Electrolyzer temperature	°C
PV(3)	Electrolyte level	L
PV(4)	Electrolyzer pressure	kPa
PV(5)	Hydrogen purity	umol/mol
PV(6)	Oxygen purity	umol/mol
PV(7)	Hydrogen separator pressure	kPa
PV(8)	Oxygen separator pressure	°C
PV(9)	Hydrogen condenser temperature	°C
PV(10)	Oxygen condenser temperature	°C
PV(11)	Oxygen electrolyte condenser temperature	°C
PV(12)	Hydrogen electrolyte condenser temperature	°C
PV(13)	Return water condenser temperature	°C
PV(14)	KOH tank level	L
PV(15)	Water tank level	L
PV(16)	Gas dryer humidity	%RH
PV(17)	Gas dryer pressure	kPa
PV(18)	Return pipeline pressure	kPa
PV(19)	Hydrogen separator electrolyte concentration	wt%
PV(20)	Oxygen separator electrolyte concentration	wt%
PV(21)	Hydrogen purifier pressure	kPa
PV(22)	Hydrogen purifier purity	umol/mol

Table 3: Summary of different type of faults in AWE system.

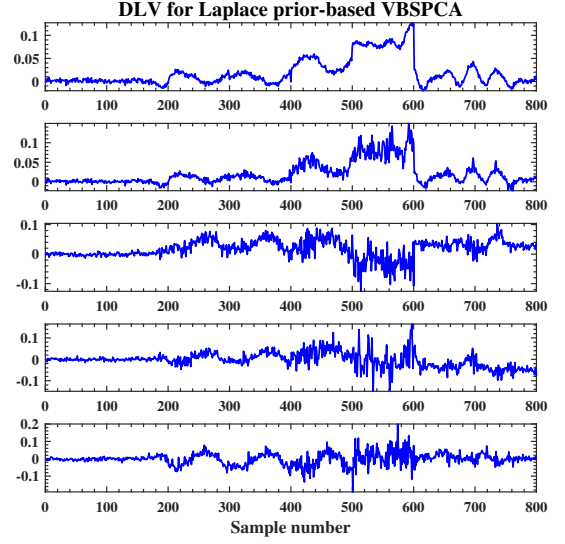
Tag	Fault description	Module
1	Electrolysis cell voltage rise	Electrolytic module
2	Electrolyte inter-electrode voltage is abnormal	Electrolytic module
3	Electrolyzer short circuit	Electrolytic module
4	Reduced gas purity	Gas-liquid separation module
5	Increased hydrogen oxygen liquid level difference	Gas-liquid separation module
6	A sharp rise or fall in the hydrogen or oxygen level	Gas-liquid separation module
7	Unstable pressure	Gas-liquid separation
8	Electrolyte stops circulating	Circulating module
9	Increased electrolyzer temperature	Circulating module
10	Water tank bubbles	Circulating module

In this section, two faults in the three modules of the hydrogen production process are analyzed. All faults in the test sample were introduced from the 201st sample. Fault 2 is the electrolyte inter-electrode voltage is abnormal. The hydrogen production rate in the alkaline water electrolyzer depends on the DC power supply output. The higher the voltage applied to the water, the faster the water molecules will split into hydrogen and oxygen gases. This is because the voltage provides enough energy to overcome the bond energy of water and drive the electrolysis reaction. When the voltage increases, more electrons are transferred to the water, making the reaction more vigorous. However, too high a voltage can also generate excessive bubbles that can lower the reaction efficiency. Therefore, it is important to choose an optimal voltage for water electrolyzer to achieve the best balance between reaction speed and effect. When the voltage is abnormal, it can cause the electrolysis reaction to not proceed properly. High voltage can cause bubbles to accumulate in the electrolyzer, posing a fire and explosion hazard. Voltage fluctuations can lead to unstable electrolysis reactions and the hydrogen production system poses an explosion risk.

By capturing the dynamic relationships of multiple variables through lagged values of each variable and regression analysis, the response of a dynamic latent variable to another variable at different points in time can be calculated. The predicted values of the latent variables calculated by the sparse VAR model are shown in Figure 2, where the five dynamic latent variables have been able to characterize the full range of changes in the system. The correlations of the obtained dynamic latent variables were checked by plotting correlation plots. The correlations among the dynamic latent variables are presented in Figure 3. As depicted, it is evident that the correlations have been eliminated through modeling, whereby the first five dynamic latent variables are extracted.

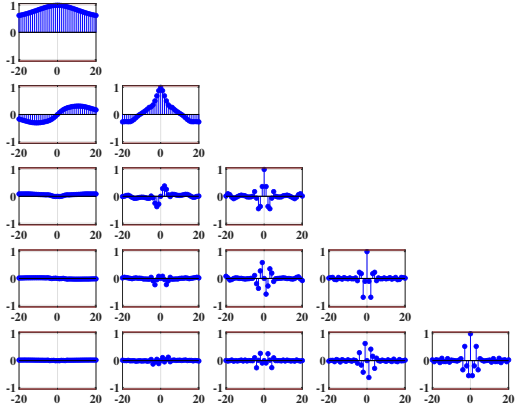


(a) Gaussian prior-based VBSPCA.

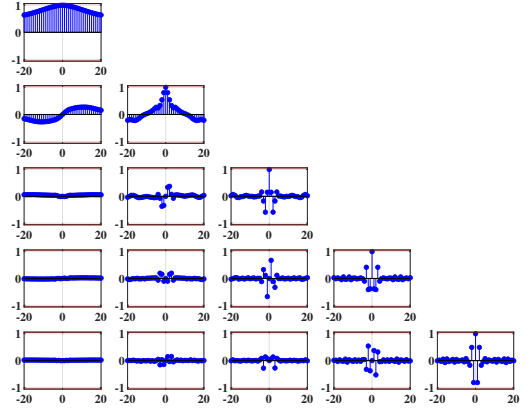


(b) Laplace prior-based VBSPCA.

Figure 2: Dynamic latent variables.



(a) Gaussian prior-based VBSPCA.

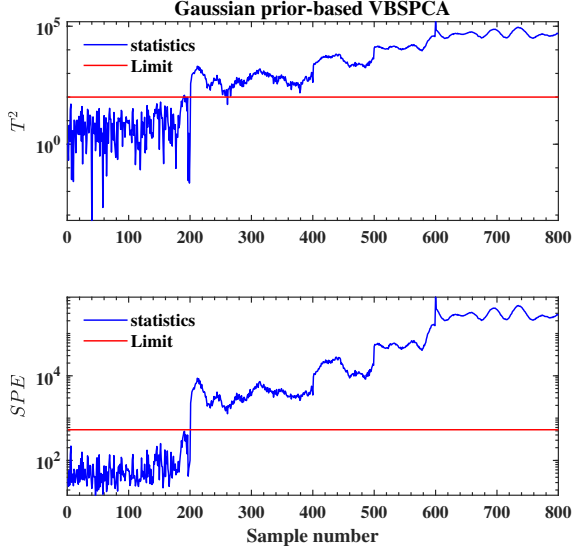


(b) Laplace prior-based VBSPCA.

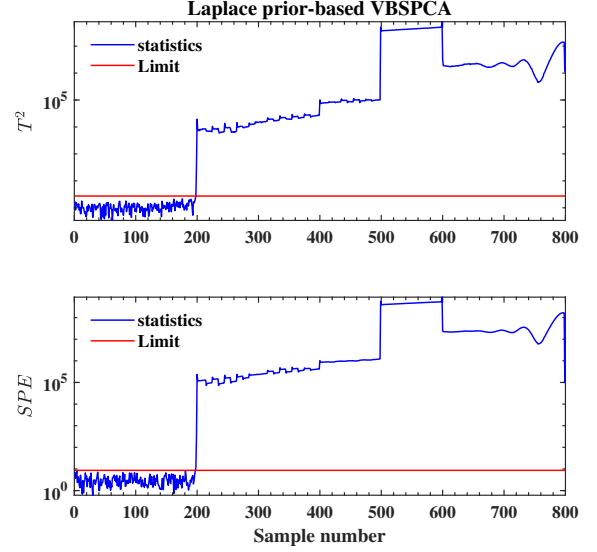
Figure 3: Autocorrelation coefficients of dynamic latent variables.

The results of the fault detection are shown in Figure 4, where the blue line represents the fault statistic, which presents the variation of the fault process. The red line is the confidence limit, and a statistic above the confidence limit indicates that a fault has been detected. In order to verify the advantages of the VBSPCA method, we compare the variational Bayesian dictionary learning (VBDL) method with the DiCCA method. The VBDL obtains sparsity by constructing a dictionary matrix, but does not consider the dynamic relationships[33]. The DiCCA method analyses the dynamic relationships of the latent variables but lacks robustness to noise perturbations. Fault 2 is heavily contaminated by noise, and three different Bayesian sparsity methods effectively detect

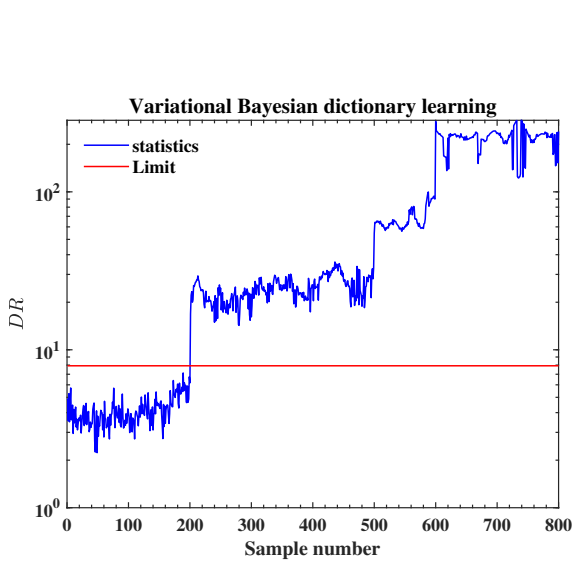
the fault, but the DiCCA method does not detect all faults. In addition, the Laplace prior-based VBSPCA method has a better fault detection ability compared to the Gaussian prior-based VBSPCA method.



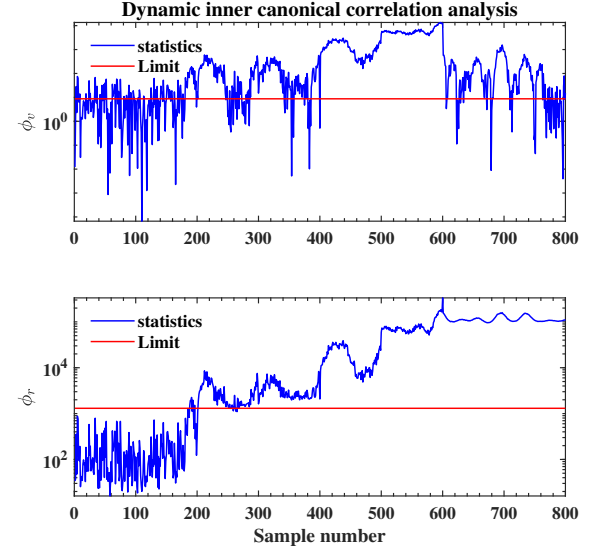
(a) Gaussian prior-based VBSPCA.



(b) Laplace prior-based VBSPCA.



(c) Variational Bayesian dictionary learning.



(d) Dynamic inner canonical correlation analysis.

Figure 4: Detection result of fault 2.

The fault diagnosis results based on  $T_2$  and  $SPE$  indices are shown in Figures 5, respectively. The horizontal axis represents the sample index, while the vertical axis represents the variable index. The RBC for each variable and sample is represented by the color depth, where deeper colors indicate higher RBC. The VBSPCA method based on both Gaussian prior and Laplace prior can diagnose the fault. The fault diagnosis results indicate that variable 24 plays a crucial role in the early stage of the fault, while variables 27 and 28 are the main factors in the later stage

of the fault. In addition, the Laplace prior-based VBSPCA suggests that the first five variables also contribute to the fault.

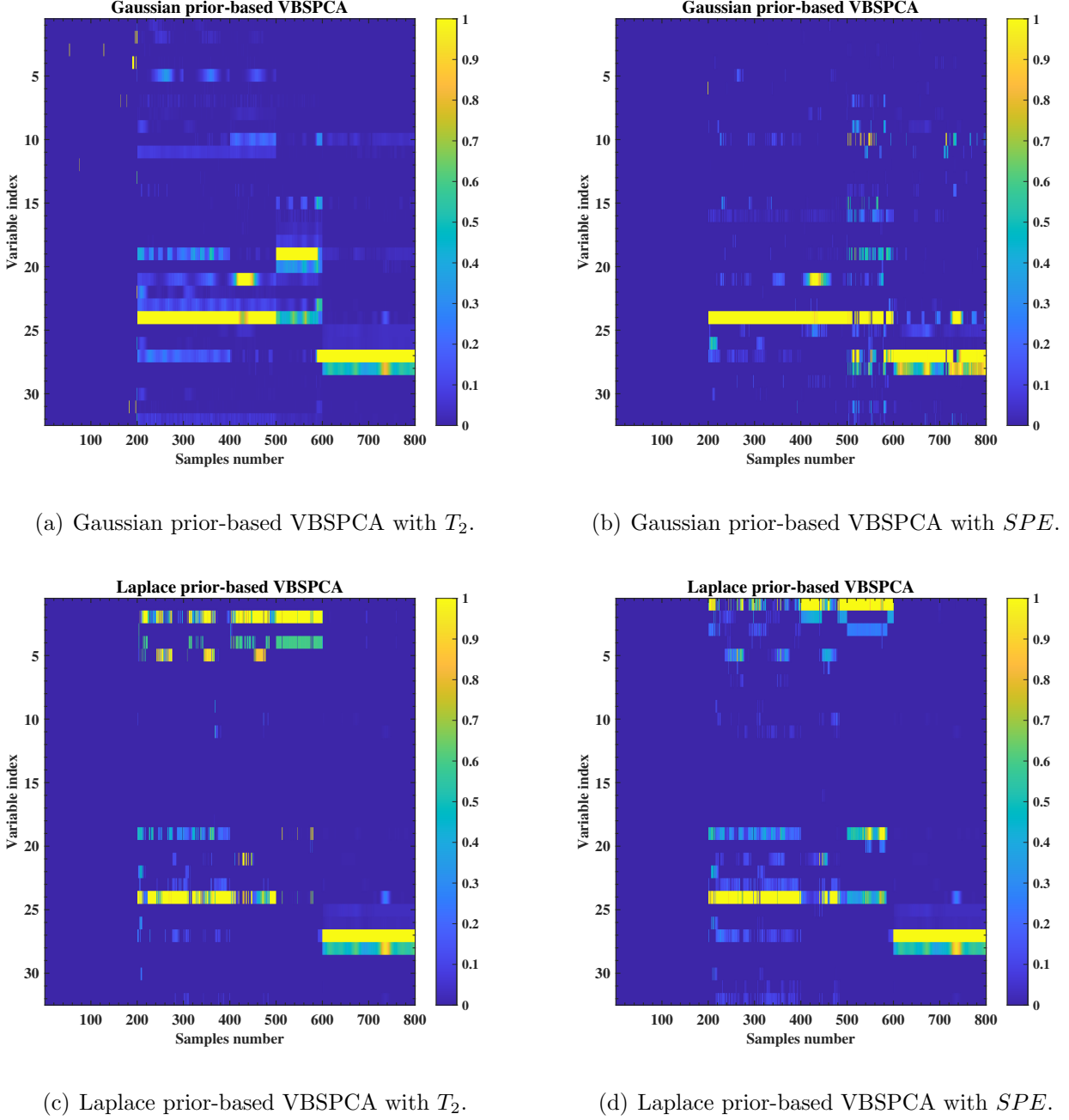
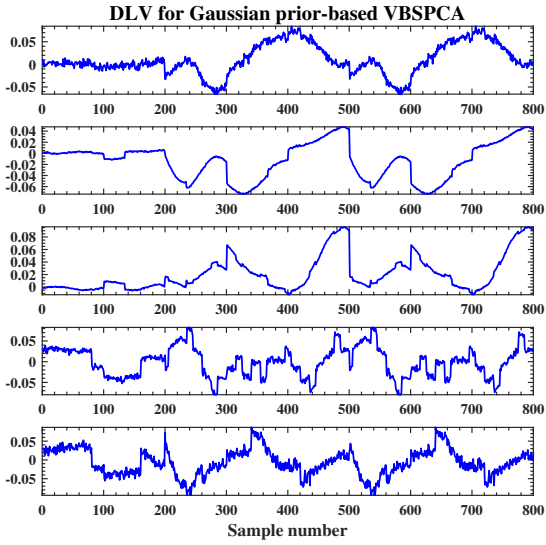


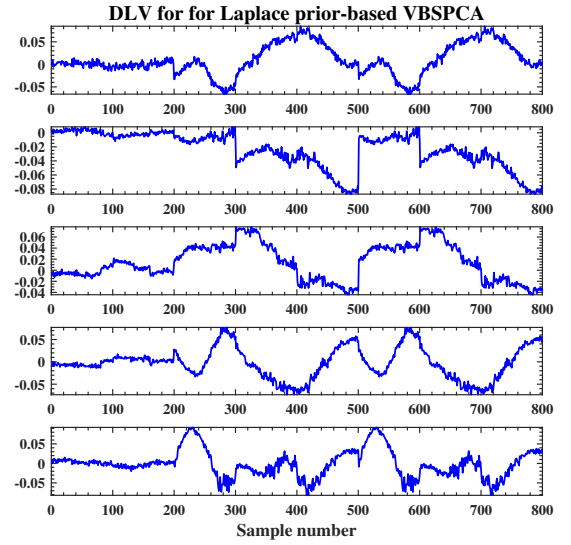
Figure 5: Diagnosis result of fault 2.

Fault 6 is a sharp rise or fall in the hydrogen or oxygen level. If there is a sudden change in the hydrogen level, it may result in hydrogen leakage, which can cause a hydrogen explosion accident, leading to casualties and equipment damage. There are several factors that can cause changes in the hydrogen level during the production process, including blockage of the balance valve or pipeline, and blockage of the outlet of the separation and defogging device. Additionally, damage to the diaphragm inside the electrolytic cell due to long-term operation can also result in improper collection of hydrogen or oxygen, leading to a sudden change in the hydrogen or oxygen liquid level.

The predicted values of the latent variables calculated by the sparse VAR model are shown in Figure 6, where the five dynamic latent variables have been able to characterize the full range of changes in the system. It is also evident from the fault detection results that the first dynamic latent variable and the detection results exhibit the highest degree of similarity. The correlations of the obtained dynamic latent variables were checked by plotting correlation plots. The correlations among the dynamic latent variables are presented in Figure 7. As depicted, it is evident that the correlations have been eliminated through modeling, whereby the first five dynamic latent variables are extracted.

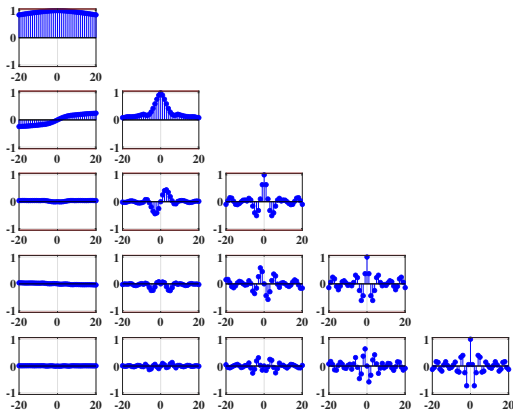


(a) Gaussian prior-based VBSPCA.

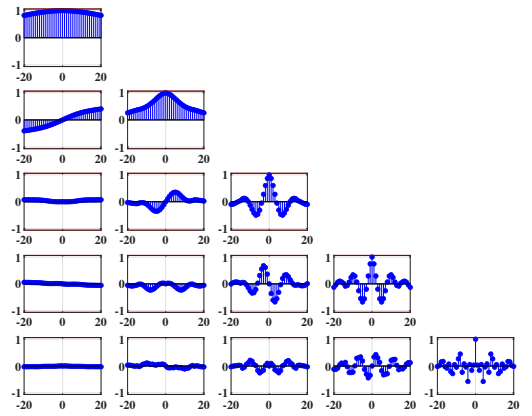


(b) Laplace prior-based VBSPCA.

Figure 6: Dynamic latent variables.



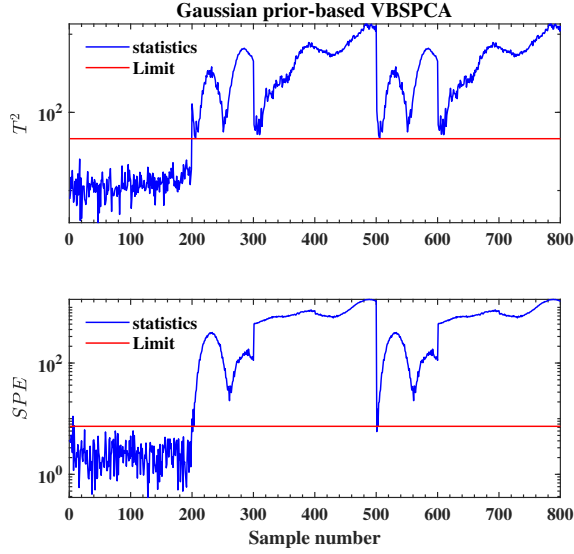
(a) Gaussian prior-based VBSPCA.



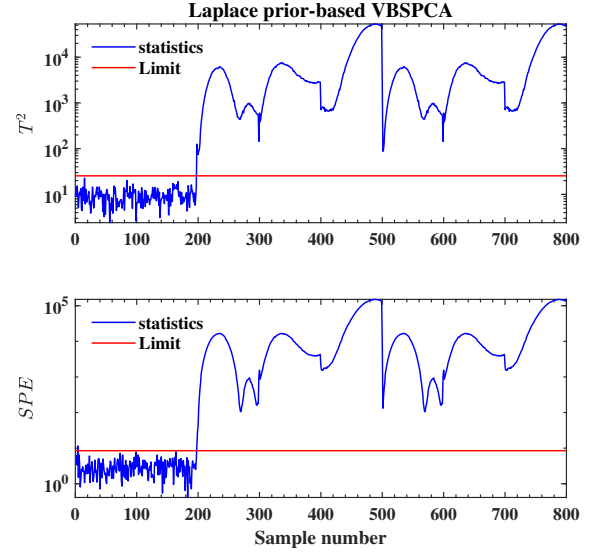
(b) Laplace prior-based VBSPCA.

Figure 7: Autocorrelation coefficients of dynamic latent variables.

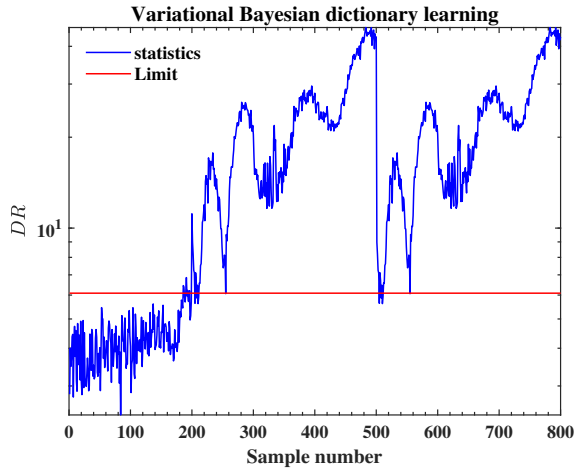
The results of the fault detection are shown in Figure 8. Fault 6 is also heavily contaminated by noise, and three different Bayesian sparsity methods effectively detect the fault, but the DiCCA method does not detect all faults. VBDL fails to obtain sufficiently good detection results. This also shows the importance of dynamic analysis for fault detection. The Laplace prior-based VBSPCA method has a better fault detection ability compared to the Gaussian prior-based VBSPCA method. This suggests that the Lasso sparsity corresponding to the Laplace prior is more favourable for feature selection and noise reduction.



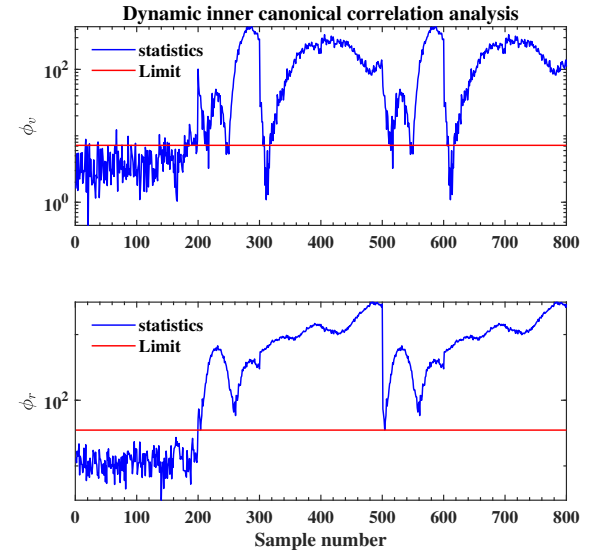
(a) Gaussian prior-based VBSPCA.



(b) Laplace prior-based VBSPCA.



(c) Variational Bayesian dictionary learning.



(d) Dynamic inner canonical correlation analysis.

Figure 8: Detection result of fault 6.

The fault diagnosis results based on  $T_2$  and  $SPE$  indices are shown in Figures 9. The fault diagnosis results indicate that in this case, fault 2 is primarily caused by variable 8, which makes an

absolute contribution. It should be noted that under different conditions, this fault may be caused by different factors. Additionally, the contribution plot based on  $SPE$  shows that variables 18 to 21 also contribute to the fault. The Gaussian prior-based VBSPCA only considers variable 8 as contributing in  $T_2$ . But the Laplace prior-based VBSPCA assumes that there are many variables that contribute to the fault, although not continuously.

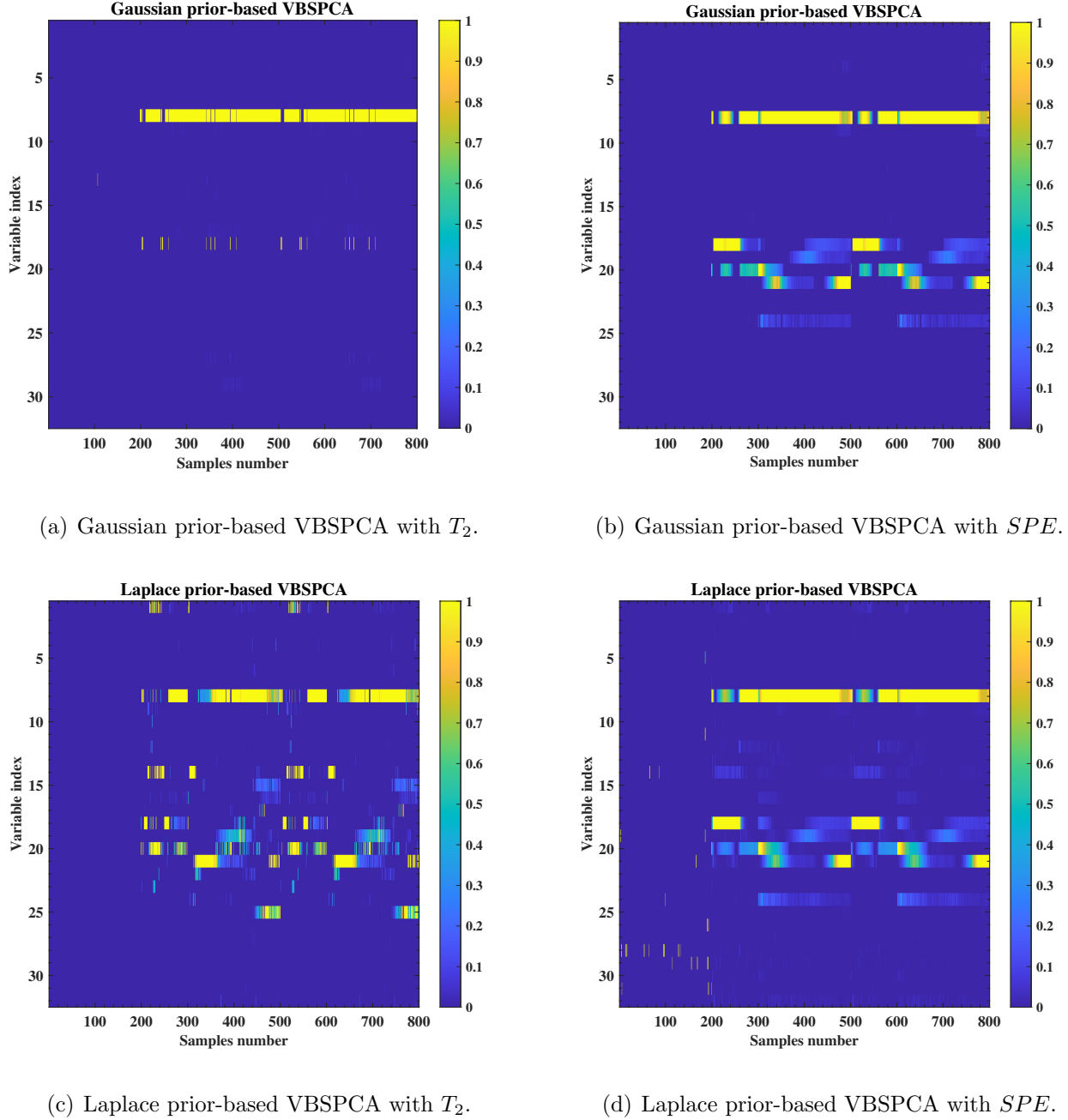


Figure 9: Diagnosis result of fault 6.

## 7. Conclusions

Hydrogen production from electrolytic water is a green method of hydrogen production, which is the main way to produce hydrogen from renewable energy such as photovoltaic and wind power.



With the rapid development of renewable energy and the update and iteration of material technology, electrolytic water hydrogen production technology has a broad development prospect. However, the safety of the hydrogen production process is often overlooked. The process monitoring of electrolytic water hydrogen production refers to the real-time measurement and control of each parameter in the electrolytic water hydrogen production system to ensure the safe, stable and efficient operation of the system. The industrial environment causes the AWE process data to be randomly contaminated by noise. In this study, VBSPCA approach is developed to detect fault of AWE industrial process. We derive two sparse PCA methods in the framework of variational inference. And it is shown that Laplace prior and Gaussian prior correspond to  $\ell_1$  regularisation and  $\ell_2$  regularisation. Constructing a sparse projection matrix is more conducive to feature selection and improves the robustness of the model to noise. The proposed method can detect faults in the production environment of industrial noise and has the capability to dynamically analyze data. The established dynamic process monitoring scheme can effectively extract the dynamic relationship of the data and obtain reliable detection and diagnosis results. Therefore, the proposed solution is attractive for green hydrogen production applications.

## References

- [1] B. M. Wise, N. B. Gallagher, The process chemometrics approach to process monitoring and fault detection, *Journal of Process Control* 6 (6) (1996) 329–348. doi:[10.1016/0959-1524\(96\)00009-1](https://doi.org/10.1016/0959-1524(96)00009-1).
- [2] R. Raveendran, H. Kodamana, B. Huang, Process monitoring using a generalized probabilistic linear latent variable model, *Automatica* 96 (2018) 73–83. doi:[10.1016/j.automatica.2018.06.029](https://doi.org/10.1016/j.automatica.2018.06.029).
- [3] J. MacGregor, T. Kourti, Statistical process control of multivariate processes, *Control Engineering Practice* 3 (3) (1995) 403–414. doi:[10.1016/0967-0661\(95\)00014-L](https://doi.org/10.1016/0967-0661(95)00014-L).
- [4] S. J. Qin, Survey on data-driven industrial process monitoring and diagnosis, *Annual Reviews in Control* 36 (2) (2012) 220–234. doi:[10.1016/j.arcontrol.2012.09.004](https://doi.org/10.1016/j.arcontrol.2012.09.004).
- [5] S. J. Qin, Y. Dong, Q. Zhu, J. Wang, Q. Liu, Bridging systems theory and data science: A unifying review of dynamic latent variable analytics and process monitoring, *Annual Reviews in Control* 50 (2020) 29–48. doi:[10.1016/j.arcontrol.2020.09.004](https://doi.org/10.1016/j.arcontrol.2020.09.004).
- [6] W. Ku, R. H. Storer, C. Georgakis, Disturbance detection and isolation by dynamic principal component analysis, *Chemometrics and Intelligent Laboratory Systems* 30 (1) (1995) 179–196. doi:[10.1016/0169-7439\(95\)00076-3](https://doi.org/10.1016/0169-7439(95)00076-3).
- [7] G. Li, S. J. Qin, D. Zhou, A new method of dynamic latent-variable modeling for process monitoring, *IEEE Transactions on Industrial Electronics* 61 (11) (2014) 6438–6445. doi:[10.1109/TIE.2014.2301761](https://doi.org/10.1109/TIE.2014.2301761).
- [8] Q. Zhang, S. Lu, L. Xie, S. Gu, H. Su, Variational Bayesian State Space Model for dynamic process fault detection, *Journal of Process Control* 124 (2023) 129–141. doi:[10.1016/j.jprocont.2023.02.004](https://doi.org/10.1016/j.jprocont.2023.02.004).
- [9] Y. Dong, S. J. Qin, A novel dynamic PCA algorithm for dynamic data modeling and process monitoring, *Journal of Process Control* 67 (2018) 1–11. doi:[10.1016/j.jprocont.2017.05.002](https://doi.org/10.1016/j.jprocont.2017.05.002).

- [10] Y. Dong, S. J. Qin, New Dynamic Predictive Monitoring Schemes Based on Dynamic Latent Variable Models, *Industrial & Engineering Chemistry Research* 59 (6) (2020) 2353–2365. doi:[10.1021/acs.iecr.9b04741](https://doi.org/10.1021/acs.iecr.9b04741).
- [11] Q. Zhang, S. Lu, L. Xie, Q. Chen, H. Su, Quality-Relevant Process Monitoring with Concurrent Locality-Preserving Dynamic Latent Variable Method, *ACS Omega* 7 (31) (2022) 27249–27262. doi:[10.1021/acsomega.2c02118](https://doi.org/10.1021/acsomega.2c02118).
- [12] X. Liu, X. Yang, Robust Variational Inference for LPV Dual-Rate Systems With Randomly Delayed Outputs, *IEEE Transactions on Instrumentation and Measurement* 70 (2021) 1–9. doi:[10.1109/TIM.2021.3067242](https://doi.org/10.1109/TIM.2021.3067242).
- [13] E. J. Candes, M. B. Wakin, An Introduction To Compressive Sampling, *IEEE Signal Processing Magazine* 25 (2) (2008) 21–30. doi:[10.1109/MSP.2007.914731](https://doi.org/10.1109/MSP.2007.914731).
- [14] B. K. Natarajan, Sparse Approximate Solutions to Linear Systems, *SIAM Journal on Computing* 24 (2) (1995) 227–234. doi:[10.1137/S0097539792240406](https://doi.org/10.1137/S0097539792240406).
- [15] R. Muthukrishnan, R. Rohini, LASSO: A feature selection technique in predictive modeling for machine learning, in: 2016 IEEE International Conference on Advances in Computer Applications (ICACA), 2016, pp. 18–20. doi:[10.1109/ICACA.2016.7887916](https://doi.org/10.1109/ICACA.2016.7887916).
- [16] R. Tibshirani, Regression Shrinkage and Selection Via the Lasso, *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1) (1996) 267–288. doi:[10.1111/j.2517-6161.1996.tb02080.x](https://doi.org/10.1111/j.2517-6161.1996.tb02080.x).
- [17] S. Ji, Y. Xue, L. Carin, Bayesian Compressive Sensing, *IEEE Transactions on Signal Processing* 56 (6) (2008) 2346–2356. doi:[10.1109/TSP.2007.914345](https://doi.org/10.1109/TSP.2007.914345).
- [18] Y. Guan, J. G. Dy, Sparse Probabilistic Principal Component Analysis.
- [19] S. D. Babacan, R. Molina, A. K. Katsaggelos, Bayesian Compressive Sensing Using Laplace Priors, *IEEE Transactions on Image Processing* 19 (1) (2010) 53–63. doi:[10.1109/TIP.2009.2032894](https://doi.org/10.1109/TIP.2009.2032894).
- [20] M. E. Tipping, Sparse Bayesian learning and the relevance vector machine, *Journal of machine learning research* 1 (Jun) (2001) 211–244.
- [21] D. Wipf, B. Rao, Sparse Bayesian learning for basis selection, *IEEE Transactions on Signal Processing* 52 (8) (2004) 2153–2164. doi:[10.1109/TSP.2004.831016](https://doi.org/10.1109/TSP.2004.831016).
- [22] C. M. Bishop, *Pattern Recognition and Machine Learning*.
- [23] N. Sammaknejad, Y. Zhao, B. Huang, A review of the Expectation Maximization algorithm in data-driven process identification, *Journal of Process Control* 73 (2019) 123–136. doi:[10.1016/j.jprocont.2018.12.010](https://doi.org/10.1016/j.jprocont.2018.12.010).
- [24] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational Inference: A Review for Statisticians, *Journal of the American Statistical Association* 112 (518) (2017) 859–877. doi:[10.1080/01621459.2017.1285773](https://doi.org/10.1080/01621459.2017.1285773).
- [25] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, L. K. Saul, An Introduction to Variational Methods for Graphical Models, *Machine Learning* 37 (2) (1999) 183–233. doi:[10.1023/A:1007665907178](https://doi.org/10.1023/A:1007665907178).
- [26] M. J. Beal, Z. Ghahramani, Variational Bayesian learning of directed graphical models with hidden variables, *Bayesian Analysis* 1 (4) (2006). doi:[10.1214/06-BA126](https://doi.org/10.1214/06-BA126).
- [27] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, University of London, University College London (United Kingdom), 2003.
- [28] S. Joe Qin, Statistical process monitoring: Basics and beyond, *Journal of Chemometrics* 17 (8-9) (2003) 480–502. doi:[10.1002/cem.800](https://doi.org/10.1002/cem.800).
- [29] M. E. Tipping, C. M. Bishop, Probabilistic Principal Component Analysis, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 61 (3) (1999) 611–622. doi:[10.1111/1467-9868.00196](https://doi.org/10.1111/1467-9868.00196).
- [30] S. W. Choi, E. B. Martin, A. J. Morris, I.-B. Lee, Fault Detection Based on a Maximum-Likelihood Principal

Component Analysis (PCA) Mixture, *Industrial & Engineering Chemistry Research* 44 (7) (2005) 2316–2327. doi:[10.1021/ie049081o](https://doi.org/10.1021/ie049081o).

- [31] M. Oppor, D. Saad (Eds.), *Advanced Mean Field Methods: Theory and Practice*, Neural Information Processing, MIT, Cambridge, Massachusetts (US), 2001.
- [32] C. F. Alcala, S. J. Qin, Reconstruction-based contribution for process monitoring, *Automatica* 45 (7) (2009) 1593–1600. doi:[10.1016/j.automatica.2009.02.027](https://doi.org/10.1016/j.automatica.2009.02.027).
- [33] Q. Zhang, L. Xie, W. Xu, H. Su, A Variational Bayesian Dictionary Learning for Process Monitoring, in: 2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS), 2023, pp. 11–15. doi:[10.1109/DDCLS58216.2023.10166191](https://doi.org/10.1109/DDCLS58216.2023.10166191).