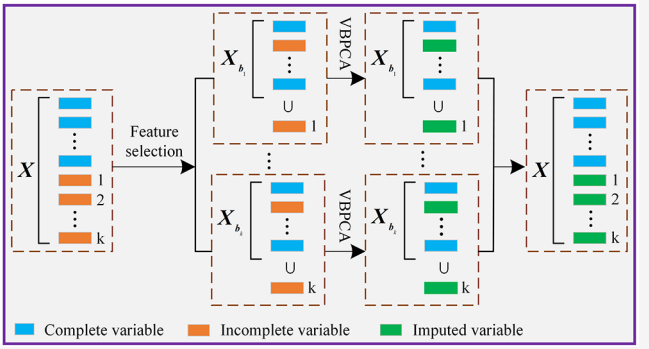
I&EC research
Industrial & Engineering Chemistry Research
pubs.acs.org/IECR
Article

Distributed Missing Values Imputation Schemes for Plant-Wide Industrial Process Using Variational Bayesian Principal Component Analysis

Linsheng Zhong, Yuqing Chang,* Fuli Wang, and Shihong Gao

Cite This: *Ind. Eng. Chem. Res.* 2022, 61, 580−593

ABSTRACT: Industrial process data often contains missing values due to network transmission errors and sensor failures, etc. Unlike some fields such as biology and climatic science, missing values imputation (MVI) for online data is necessary for industrial processes, because most of the data-based intelligent decision support systems demand a complete training data set and online samples. To the best of the authors' knowledge to date, limited results on MVI for both a training data set and online samples are reported. How to achieve a comprehensive and high-precision MVI scheme in line with industrial reality is still an open problem. Directed toward the complicated correlations among variables of plant-wide industrial process (PWIP), we first propose an improved feature subset selection algorithm based on the time shift correlation and a newly defined selection criterion. Second, for each variable with missing values, the feature subset with strong correlations to this variable is selected. In this way, the time-lagged correlations both within and across variables are made full use of. Through applying variational Bayesian principal component analysis (VBPCA) into the resulting feature subsets, a novel distributed shift correlation-based VBPCA (DSCVBPCA) technique is developed to achieve a better imputation effect. Thereafter, the moving window strategy and the modified DSCVBPCA with tactfully set parameters are integrated to accomplish the MVI for online samples. Finally, the experiments much closer to the actual situations of PWIP are conducted on a numerical example and gold hydrometallurgy, indicating that the proposed imputation scheme can be a promising alternative for the MVI of industrial processes.



1. INTRODUCTION

Due to network transmission errors, sensor faults or multiple sampling rates, etc., the loss of measured data or observations at some unexpected moments frequently occurs in actual production processes.[1] Since high-quality enterprise-level intelligent decision support systems[2,3] (e.g., data-driven process monitoring) are highly dependent on high-quality complete process data, the missing values (MV) processing technology has become crucial for industrial processes.[4] Among various MV treatment schemes, the MV imputation (MVI) which replaces the MV with the plausible values derived from statistics or machine learning algorithms is popular for restoring incomplete data sets. This work is devoted to developing an effective MVI method for the plant-wide industrial process (PWIP) that is usually equipped with multi-interconnected running units and distributed control systems. Different from the fields such as biology and climatic science, the measured data of the PWIP is characterized by information redundancy, data collinearity, and coexistence of static and dynamic correlations,[5,6] and therefore, researching the MVI algorithm suitable for the PWIP is of great challenge and fundamental importance.

The main techniques used for disposing MV of industrial processes include deletion,[7] robust model[8−10] and MVI.[11−16] The deletion directly ignores the samples with MV, which is simple but leads to the loss of valuable information on incomplete samples. When the number of samples with MV is large, the data analysis result tends to be unreliable. In fact, it is proved that the proper utilization of the samples with MV helps to obtain the data distribution and reduce the estimation bias caused by only using complete samples.[17] Given this, based on the available information on incomplete data sets, the robust model estimates the parameters of the desired model by employing the suitable parameter identification method (e.g.,

Received: September 26, 2021
Revised: December 1, 2021
Accepted: December 1, 2021
Published: December 27, 2021

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

the maximum likelihood estimation). However, the portability of robust models is not flexible enough. The MVI restores the incomplete data sets by replacing the MV with the plausible values derived from statistics or machine learning algorithms. Since most of the data-driven intelligent decision support systems are developed on complete data sets, the MVI technique is more universal. Recently, a number of MVI algorithms have been presented.[18−22] The mean imputation uses the average value of a variable's observations to impute the variable's MV, but it results in a smaller variance.[18] The k nearest-neighbor (kNN) imputation is to find out the k nearest-neighbor samples from the data set except for the samples that have MV at the same positions as those to be imputed.[19] Based on the modified kNN, Sahri et al.[11] propose an iterative imputation strategy and apply it to the incomplete dissolved gas analysis data set. However, the kNN-based imputation neglects the correlation between variables. In contrast, the expectation maximization (EM) imputation relies on the iterated analysis of the linear regression between the variables with MV and other variables with available values.[21] The EM algorithm is widely used owing to its satisfactory imputation effect; for instance, Wang et al.[12] establish a fault detection model on the basis of the EM-imputed data set. It is noteworthy that since the linear regression model is based on the original variables, the collinearity between variables will cause the instability of model parameter estimation or even lead to the divergence of the algorithm. Different from the EM imputation, the principal component (PC) analysis (PCA) transforms the original variables into a few PCs that are the linear combinations of original variables. For the regression model built on PCs, the collinearity is avoided without losing important data features. As a maximum-likelihood reformulation of PCA, the probability PCA (PPCA) is capable of dealing with MV in data sets,[22] but the issue of overfitting becomes severe when faced with high-dimensional and sparsely distributed data. On account of this, the variational Bayesian learning is integrated into PPCA to form the VBPCA that can effectively restrain the overfitting and achieve the MVI with higher accuracy.[22,23] With VBPCA imputing the missing data, a key quality related indices prediction model is constructed.[14] Based on the above analysis, VBPCA is selected to construct our MVI algorithm. Lately, some scholars have explored the deep neural networks (DNN) based MVI,[15,16] but the complex parameter adjustment makes these solutions suffer from long training processes.[24]

The above-mentioned achievements[11−16] have provided excellent MVI performance in the industrial processes; however, they all belong to the centralized models without selected features/variables. For the PWIP characterized by long processes and multiple procedures, the centralized models may not be the most suitable because they may involve the variables that are irrelevant or have little correlations with the variables that need to be imputed. In response to this, the distributed modeling strategy aiming at the large-scale industrial process is explored.[25] In the framework of distributed modeling, the whole feature set is first decomposed into multiple disjoint or overlapping subsets by the domain knowledge-based method[26,27] or data-based method,[28,29] which is called feature subset selection, and then the imputation algorithm is applied to each feature subset. Actually, feature subset selection is the process of identifying and removing the irrelevant features as much as possible from a training data set, which can reduce the data dimension and improve the model performance. Jiang et al.[27] develop a distributed MVI scheme named neighborhood VBPCA (NVBPCA) for the industrial tail gas treatment process, where the variables in the same subprocesses with hardware connections are grouped into the same feature subset. Sefidian et al.[28] adopt a local mutual information-based feature subset selection to identify the features that have high correlations, and then apply the regression models into the feature subsets to achieve better imputation effect. However, the aforesaid results[26−29] only consider the static characteristics of data but disregard the dynamic characteristics. The dynamic characteristics of PWIP data are reflected in that the interaction between variables usually takes time, meaning that the cross-correlations with nonzero lags between variables may be greater than those with zero lags.[30] In view of this, we propose an improved feature subset selection algorithm by employing the time shift correlation and defining a new selection criterion, so that the feature subset with strong correlations to each variable with MV is selected and the time-lagged correlations both within and across variables are made full use of. Thereafter, VBPCA can be applied into the resulting feature subsets, and a new distributed shift correlation-based VBPCA (DSCVBPCA) MVI scheme is developed.

Intelligent decision support systems for industrial processes usually include offline modeling and online application because the industrial process data belong to the streaming data that is continuously generated by different sources.[31] The offline modeling is based on the historical data set that meets the corresponding quality index (this type of historical data set is also called training data set), while the online application is to input the online samples into the offline developed model to detect whether there is an abnormal state in the process. Therefore, the MVI for online samples is another important research topic for industrial processes. In terms of the online samples MVI, the foregoing achievements either directly neglect the incomplete samples[11−13,27,28] or require a sufficient number of complete samples in the training data set.[14−16] However, acquiring a sufficient number of complete samples becomes intractable when the MV are part of the inherent structure of the study object.[32] To the best of the authors' knowledge, during 2006 to 2019, there is almost no study on the MVI that has considered both the training data set and online samples simultaneously when performing the missing data experiments for a specific missing rate.[33] In 2020, Yu and Zhao[34] investigated the MVI algorithm disposing both the training data set and online samples by means of the low-rank matrix completion (LRMC);[20] however, this solution lacks necessary modifications that should be made according to the data characteristics of industrial processes. The existing MVI algorithms such as mean, kNN, and the regression models considering different value-missing cases may be used for online samples MVI, but the performance of offline developed models is difficult to maintain in online applications because the process states tend to change with parameter drift.[35] To adapt the parameter drift with a maximal extent, the online update strategies such as the moving window (MW)[35] and just-in-time (JIT)[36] model are commonly used. Since the JIT model may disrupt the sequence correlations of samples, we adopt the MW model that receives all recent samples. By combining the MW model and the modified DSCVBPCA, a suitable online samples MVI scheme for PWIP is constructed. The main contributions of this paper are summarized as follows:

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

(1) Directed toward the complicated correlations among the variables of PWIP, an enhanced feature subset selection algorithm is proposed by employing the time shift correlation and defining a new selection criterion, so that the feature subset with strong correlations to each variable with MV is selected. By applying VBPCA to the resulting feature subsets, a better imputation effect is achieved.

(2) The existing MVI achievements for industrial processes either neglect the MVI of online samples[11−13,27,28] or require a sufficient number of complete samples in the training data set.[14−16] To the best of the authors' knowledge to date, the reported literature on MVI for both training data set and online samples is scarce. On account of this, we first propose a novel DSCVBPCA technique for the MVI of the training data set and then apply the MW model and the modified DSCVBPCA to the online samples MVI.

(3) The designed experiments on the numerical example and gold hydrometallurgy are much closer to the actual situations of industrial processes. The effectiveness and practicability demonstrate that the developed imputation scheme can be a promising candidate for the MVI of PWIP.

The rest is organized as follows. The preliminaries of VBPCA are introduced in Section 2. The proposed MVI algorithm DSCVBPCA is elaborated in Section 3. Section 4 presents the experimental results of numerical example and gold hydrometallurgy process, and the paper ends with the conclusion in Section 5.

2. MISSING VALUES IMPUTATION VIA VARIATIONAL BAYESIAN PRINCIPAL COMPONENT ANALYSIS

The VBPCA is a modified version of PPCA.[22] Three elementary processes constitute the VBPCA imputation algorithm: PC regression, VB estimation, and EM-like repetitive algorithm. Given an incomplete data set $\boldsymbol{X} \in R^{d\times n}$, where d and n denote the number of variables and samples, respectively. Then the PPCA model is expressed as

$$x_{ij} = \boldsymbol{w}_i^T \boldsymbol{t}_j + m_i + \varepsilon_{ij} \tag{1}$$

where $\boldsymbol{w}_i \in R^{c\times 1}$ is the projection vector, $\boldsymbol{t}_j \in R^{c\times 1}$ is the score vector, c denotes the number of PCs, m_i denotes the bias, and ε_{ij} denotes the noise. Assume that the distributions of $\boldsymbol{t}_j$ and ε_{ij} are Gaussian as follows:

$$p(\boldsymbol{t}_j) = N(\boldsymbol{t}_j|0, \boldsymbol{I}) \tag{2}$$

$$p(\varepsilon_{ij}) = N(\varepsilon_{ij}|0, v_x) \tag{3}$$

Model parameters $\boldsymbol{w}_i$, m_i, v_x and $\boldsymbol{t}_j$ can be identified by EM algorithm. In view of the overfitting problem encountered by traditional PPCA, a common approach is to penalize the parameter values corresponding to more complicated explanations of data. In the Bayesian formulation, an equivalent treatment is to introduce prior distributions over model parameters:[23]

$$p(\boldsymbol{m}) = N(\boldsymbol{m}|0, v_m\boldsymbol{I}) \tag{4}$$

$$p(\boldsymbol{W}) = \prod_{k=1}^{c} N(\boldsymbol{W}_{:k}|0, v_{w,k}\boldsymbol{I}) \tag{5}$$

During variational approximation, the hyperparameters v_m and $v_{w,k}$ can be updated. If the relevance evidence of the k-th PC is weak for reliable data modeling, the corresponding $v_{w,k}$ tends to zero. In this way, the automatic selection of the right number of PCs required by PCA is allowed, which is known as automatic relevance determination. Suppose that the maximum-likelihood estimation of hyperparameters $\xi = (v_x, v_{w,k}, v_m)$ is performed on the probabilistic model defined by (1)−(5), and this can be implemented via EM algorithm if treating model parameters $\theta = (\boldsymbol{W}, \boldsymbol{T}, \boldsymbol{m})$ as hidden variables. Implementation of the EM algorithm needs calculating the posterior $p(\theta|\boldsymbol{X},\xi)$ of hidden variables on the E-step. However, the true posterior $p(\theta|\boldsymbol{X},\xi)$ has no analytic form and one possible solution is to approximate it with a simpler probability density function $q(\theta)$. $q(\theta)$ is commonly formulated as the mean-field variational family. By means of variational approach, the E-step is revised to update the approximation $q(\theta)$ so that the cost function below is minimized

$$\begin{aligned} C(q(\theta), \xi) &= \int q(\theta)\log\frac{q(\theta)}{p(\boldsymbol{X}, \theta|\xi)}\mathrm{d}\theta \\ &= \int q(\theta)\log\frac{q(\theta)}{p(\theta|\boldsymbol{X},\xi)}\mathrm{d}\theta - \log p(\boldsymbol{X}|\xi) \end{aligned} \tag{6}$$

where $\int q(\theta)\ \log(q(\theta)/p(\theta|\boldsymbol{X},\xi))\ \mathrm{d}\theta$ is the Kullback−Leibler divergence between the true posterior and its approximation. On the M-step, the approximation $q(\theta)$ is utilized as it was the true posterior $p(\theta|\mathbf{X},\xi)$ for the sake of increasing the likelihood $p(\mathbf{X}|\xi)$. We can regard this as the minimization of (6) relative to ξ. Next, the optimization procedure can be executed to update the hyperparameters and model parameters iteratively. The updating rules are given as follows.

The updating rule of PCs:

$$\Sigma_{\boldsymbol{t}_j} = v_x\left[v_x\boldsymbol{I} + \sum_{i\in O_j}\left(\bar{\boldsymbol{w}}_i\bar{\boldsymbol{w}}_i^T + \Sigma_{\boldsymbol{w}_i}\right)\right]^{-1} \tag{7}$$

$$\bar{\boldsymbol{t}}_j = \frac{1}{v_x}\Sigma_{\boldsymbol{t}_j}\sum_{i\in O_j}\bar{\boldsymbol{w}}_i(x_{ij} - \bar{m}_i) \tag{8}$$

where $\bar{\boldsymbol{w}}_i$, $\bar{\boldsymbol{t}}_j$ and $\bar{m}_i$ are the posterior means of $\boldsymbol{w}_i$, $\boldsymbol{t}_j$ and m_i, respectively; $\Sigma_{\boldsymbol{w}_i}$ and $\Sigma_{\boldsymbol{t}_j}$ are the posterior covariances of $\boldsymbol{w}_i$ and $\boldsymbol{t}_j$, respectively; O_j is the set of i indices whose corresponding x_{ij} is available. The updating rule of the bias and the matrix $\boldsymbol{W}$:

$$\bar{m}_i = \frac{v_m}{|O_i|(v_m + v_x/|O_i|)}\sum_{j\in O_i}(x_{ij} - \bar{\boldsymbol{w}}_i^T\bar{\boldsymbol{t}}_j) \tag{9}$$

$$\tilde{m}_i = \frac{v_x v_m}{|O_i|(v_m + v_x/|O_i|)} \tag{10}$$

$$\Sigma_{\boldsymbol{w}_i} = v_x\left[v_x diag(v_{w,k}^{-1}) + \sum_{j\in O_i}\left(\bar{\boldsymbol{t}}_j\bar{\boldsymbol{t}}_j^T + \Sigma_{\boldsymbol{x}_j}\right)\right]^{-1} \tag{11}$$

$$\bar{\boldsymbol{w}}_i = \frac{1}{v_x}\Sigma_{\boldsymbol{w}_i}\sum_{j\in O_i}\bar{\boldsymbol{t}}_j(x_{ij} - \bar{m}_i) \tag{12}$$

where O_i is the set of j indices whose corresponding x_{ij} is available, $|O_i|$ is the number of elements of O_i, and $\tilde{m}_i$ is the posterior variance of m_i. Parameters of variances are given as

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

$$v_x = \frac{1}{N}\sum_{ij\in O}\left[(x_{ij} - \bar{\boldsymbol{w}}_i^T\bar{\boldsymbol{t}}_j - \bar{m}_i)^2 + \tilde{m}_i + \bar{\boldsymbol{w}}_i^T\Sigma_{\boldsymbol{t}_j}\bar{\boldsymbol{w}}_i + \bar{\boldsymbol{t}}_j^T\Sigma_{\boldsymbol{w}_i}\bar{\boldsymbol{t}}_j + \mathrm{tr}\left(\Sigma_{\boldsymbol{t}_j}\Sigma_{\boldsymbol{w}_i}\right)\right] \quad (13)$$

$$v_{w,k} = \frac{1}{d}\sum_{i=1}^{d}(\bar{w}_{ik}^2 + \tilde{w}_{ik}) \quad (14)$$

$$v_m = \frac{1}{d}\sum_{i=1}^{d}(\bar{m}_i^2 + \tilde{m}_i) \quad (15)$$

where $\tilde{w}_{ik}$ is the k-th element on the diagonal of $\Sigma_{\boldsymbol{w}_i}$.

The MV of incomplete data set $\boldsymbol{X} \in R^{d\times n}$ can be imputed by the following formula:

$$\hat{x}_{ij} = \bar{\boldsymbol{w}}_i^T\bar{\boldsymbol{t}}_j + \bar{m}_i \quad (16)$$

3. DESIGN OF DSCVBPCA IMPUTATION ALGORITHM

3.1. Problem Statement and Motivation. The closed-loop control strategy is widely used in disturbance suppression, production safety guarantee, and profit maximization of industrial processes.[37] The controller regulates process variables' behaviors in an intricate manner. In the presence of external disturbances, the variables under control will evolve around the steady state and show some degree of autocorrelation characteristics.[30] Fluctuations of process variables caused by external disturbances are depicted in Figure 1. Because the modern industrial production usually has long processes, there is a time lag in the impact of one subprocess on another. On account of this, it is pointed out in ref 38 that some process variables affect others with time lags. It means that the cross-correlations with nonzero lags between variables may be greater than those with zero lags, which endows the process data of practical productions with dynamic characteristics.

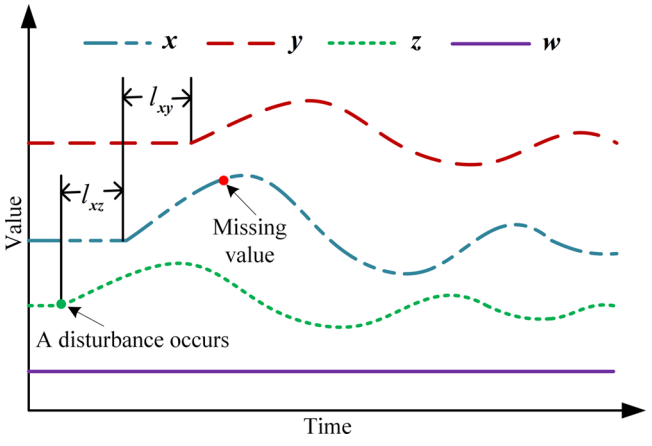


Figure 1. Fluctuations of process variables caused by external disturbances.

When value missing occurs, one should first decide what type of information can be used for MVI. Given a training data set, the past, current and future information on variables can be obtained concurrently except for the start and end stages. As shown in Figure 1, the missing value x_t (red dot) of variable $\boldsymbol{x}$ is not only related to the past information $z_{t-l_{xz}}$, but also related to the future information $y_{t+l_{xy}}$. Moreover, x_t is partially determined by other variables, and obviously, the variable $\boldsymbol{x}$ also has a certain degree of autocorrelations; that is, successive observations are not independent and x_t is also related to its own past and future information. The autocorrelation that helps to depict the evolution of a process over time is very important.[39] However, only considering the above information is not comprehensive. There may exist cross-correlations with zero lags between variables, namely, spatial correlations. For example, for two rotating equipment with hardware synchronization, if there is no obvious equipment wear, the speed ratio has nothing to do with time. In a word, static and dynamic characteristics usually coexist in actual production processes.[30]

The MVI of online samples is different from that of the training data set, mainly because the online data are coming one after another, and the primary task is to impute the MV of current online samples, whereas it is not necessary to impute the MV of past online data. Besides, the future information on current online samples cannot be obtained when imputing the MV of current online samples. In Figure 1, the missing value of the current online sample of variable $\boldsymbol{x}$ is x_t (red dot). Although x_t is related to the future information $y_{t+l_{xy}}$, $y_{t+l_{xy}}$ is unknown at this time. Therefore, only current and past information related to x_t can be used to complete online MVI.

It is notable that when a disturbance occurs, the variable $\boldsymbol{w}$ does not change. Therefore, the variable $\boldsymbol{w}$ cannot provide useful information for the imputation of the current missing value x_t, so $\boldsymbol{w}$ should be removed when establishing the imputation model for x_t. In other words, during missing value imputation, features that are more relevant to missing values should be retained whereas those that are irrelevant or less relevant to missing values should be removed. For the PWIP characterized by long processes and multiple procedures, the centralized models may involve the variables that are irrelevant or have little correlations with the variables that need to be imputed. Hence, the distributed modeling strategy may be more suitable for PWIP. In the framework of the distributed MVI scheme, as shown by Figure 2, the whole feature set is first decomposed into multiple disjoint or overlapping subsets, which is called feature subset selection, and then the imputation algorithm is applied to each feature subset. Numerous studies demonstrate that a suitable feature selection algorithm can not only reduce the data dimension but also improve the model performance.[25−29]

3.2. Feature Subset Selection Based on Time Shift Correlation. As mentioned previously, the interrelationships among measured variables can be described more accurately if different time lags are taken into account. Inspired by this, a new feature subset selection algorithm based on time shift correlation is proposed herein. Given an incomplete data set $\boldsymbol{X} \in R^{d\times n}$ with d and n respectively denoting the number of variables and samples, $\boldsymbol{x}_i$ represents the variables (rows of $\boldsymbol{X}$) with MV, $\Omega(i \in \Omega)$ represents the index set of i corresponding to the variable with MV, $l \in [-f, f]$ denotes the time shift coefficient, and $j = 1, 2, \cdots, d$. Since some values of $\boldsymbol{x}_{i,t}$ and $\boldsymbol{x}_{j,t+l}$ are missing, we only use the instances in which observations of $\boldsymbol{x}_{i,t}$ and $\boldsymbol{x}_{j,t+l}$ are both existing. Calculation formulas are given by (17)−(19):

$$r_i(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{j,t+l}) = \left|\frac{\mathrm{Cov}(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{j,t+l})}{\sqrt{\mathrm{Var}(\boldsymbol{x}_{i,t})\mathrm{Var}(\boldsymbol{x}_{j,t+l})}}\right| \quad (17)$$

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

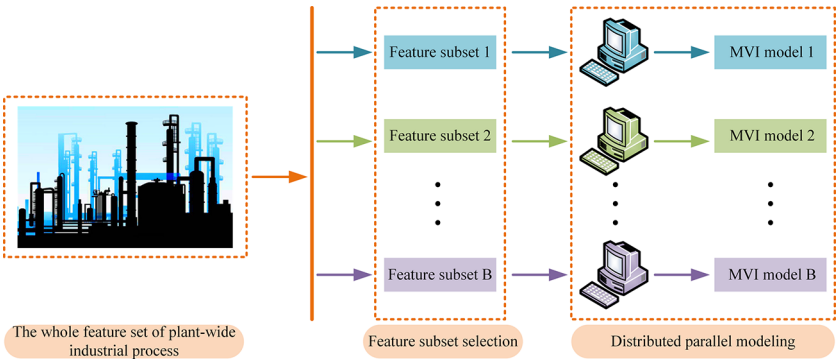


Figure 2. Distributed MVI scheme for PWIP.

$$\boldsymbol{x}_{i,t} = \begin{cases} [x_{i,1}, x_{i,2}, \ldots, x_{i,n-l}] \in R^{1\times(n-l)}, & l \geq 0 \\ [x_{i,1-l}, x_{i,2-l}, \ldots, x_{i,n}] \in R^{1\times(n+l)}, & l < 0 \end{cases} \tag{18}$$

$$\boldsymbol{x}_{j,t+l} = \begin{cases} [x_{j,1+l}, x_{j,2+l}, \ldots, x_{j,n}] \in R^{1\times(n-l)}, & l \geq 0 \\ [x_{j,1}, x_{j,2}, \ldots, x_{j,n+l}] \in R^{1\times(n+l)}, & l < 0 \end{cases} \tag{19}$$

Note that $r_i(\boldsymbol{x}_{i,t}, \boldsymbol{x}_{j,t+l}) = 1$ when $i = j$ and $l = 0$, and thus the dimension of $\boldsymbol{r}_i$ can be reduced to $\boldsymbol{r}_i \in R^{1\times[d(2f+1)-1]}$. This is a direct indicator used to measure the correlations among variables.[40] The maximum time shift coefficient f is chosen in accordance with the characteristics of actual processes and sampling intervals. If the actual time lags are relatively long or the sampling intervals are relatively short, a larger f should be selected; otherwise, a smaller f is suitable. After calculating vector $\boldsymbol{r}_i$ for the i-th variable, feature subset selection is accomplished via selecting variables with different time shift coefficients. In general, features with high correlation should be chosen to form the corresponding feature subset to impute the i-th variable more precisely. To this end, first, sort the elements of $\boldsymbol{r}_i$ in descending order, i.e., $\boldsymbol{r}_i = [r_{i,1}, r_{i,2}, \cdots, r_{i,d(2f+1)-1}]$ with $r_{i,1} \geq r_{i,2} \geq \cdots \geq r_{i,d(2f+1)-1}$. Second, based on the vector $\boldsymbol{r}_i$, we define a feature subset selection criterion (20) named cumulative percentage correlation (CPC):

$$CPC_i = \frac{\sum_{g=1}^{k} r_{i,g}}{\sum_{g=1}^{d(2f+1)-1} r_{i,g}} \geq \alpha \tag{20}$$

where the cutoff value α varies within $(0,1]$, making it easier to analyze the influence of α with different values on the imputation performance through experiments. Formula 20 explicitly indicates that the first k dominant features should be retained in the feature subset corresponding to the i-th variable. Meanwhile, to avoid selecting the features with less correlation into the subset, an auxiliary condition is added as follows:

$$r_{i,k} > 0.2 \tag{21}$$

That is, when $r_{i,k+1} \leq 0.2$, stop adding features to the feature subset. Generally speaking, features that are unrelated or weakly related to MV may be detrimental to MVI.

Remark 1. Since the PWIP is usually characterized by long processes and multiple procedures, the degree of linear dependence between process variables is subject to large variation. Therefore, the specificity of each variable should be taken into account when designing the feature subset selection algorithm. Theoretically, if only Formula 20 is used for feature selection, the α corresponding to the best feature subset of each variable is likely to be different. When there are many variables, finding the α suitable for all feature subsets becomes quite cumbersome. In view of this, we use the same α when selecting the feature subset for each variable. However, when the correlations between variables vary greatly, features with little correlation may be selected into the subset, which may increase the dimension of modeling data or reduce the imputation accuracy. On account of this, we add an auxiliary condition represented by Formula 21. By combining Formula 20 and Formula 21, the proposed method obtains a satisfactory imputation performance while maintaining certain flexibility and simplicity.

3.3. Missing Values Imputation of Training Data set.

Given an incomplete data set $\boldsymbol{X} \in R^{d\times n}$, the goal of MVI is to restore MV as accurately as possible by feat of the internal relationships among observations and MV. Obviously, the past, current and future information on variables can be obtained simultaneously except for the start and end stages. If all information related to MV is fully utilized, it is likely to improve the imputation accuracy.

Input: An incomplete data set $\boldsymbol{X} \in R^{d\times n}$, the maximum time shift coefficient f, and the cutoff value α of CPC.

Step 1: Calculate the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ of $\boldsymbol{X}$ using the observed data, and convert $\boldsymbol{X}$ to a matrix with zero mean and unit variance, so that the influence of variables' amplitude can be eliminated and the convergence rate of the algorithm can be accelerated. Construct the time shift duplicate matrix[30] $\boldsymbol{X}_a = [\boldsymbol{X}_{t-f}^T, \cdots, \boldsymbol{X}_{t-1}^T, \boldsymbol{X}_t^T, \boldsymbol{X}_{t+1}^T, \cdots, \boldsymbol{X}_{t+f}^T]^T \in R^{d(2f+1)\times(n-2f)}$ for $\boldsymbol{X}$.

Step 2: Traverse all the rows of $\boldsymbol{X}_t$ and record the row coordinates (variable index) where MV exist to Ω.

Step 3: For each $i \in \Omega$, calculate its time shift correlation coefficient $\boldsymbol{r}_i = [r_{i,1}, r_{i,2}, \cdots, r_{i,d(2f+1)-1}]$ by (17) and sort the elements of $\boldsymbol{r}_i$ in descending order. Based on (20)−(21), select the feature subset $\boldsymbol{b}_i$ with stronger correlation with $\boldsymbol{x}_{i,t}$ to get the best subset $\boldsymbol{X}_i = [\boldsymbol{x}_{i,t}^T, \boldsymbol{X}_{b_i}^T]^T$ to impute $\boldsymbol{x}_{i,t}$. $\boldsymbol{x}_{i,t}$ denotes the i-th row of $\boldsymbol{X}_t$. Note that $\boldsymbol{X}_{b_i}$ is formed by selecting the rows of $\boldsymbol{X}_a$ that correspond to the row coordinates recorded in $\boldsymbol{b}_i$.

Step 4: For each $i \in \Omega$, use VBPCA to impute $\boldsymbol{X}_i$, and then the complete observation vector $\boldsymbol{x}_{i,t}^{com}$ corresponding to $\boldsymbol{x}_{i,t}$ can be obtained. Afterward, replace the i-th row of $\boldsymbol{X}_t$ with $\boldsymbol{x}_{i,t}^{com}$ to get the complete matrix $\boldsymbol{X}_t^{com}$.

Step 5: When calculating the shift correlation coefficient of variables, the number of columns (samples) of $\boldsymbol{X}_t$ becomes $(n-2f)$ due to the time shift duplication of $\boldsymbol{X}$. That is, the columns 1 to f and (end-f+1) to end of $\boldsymbol{X}$ have not been imputed by

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

Step 4. If there are MVs in these columns, they need to be specialized. In this work, VBPCA is adopted to impute $\boldsymbol{X}$, and the columns 1 to f and (end-f+1) to end of the imputed matrix are added to $\boldsymbol{X}_t^{com}$, so that the matrix $\boldsymbol{X}^{com}$ with the same number of rows and columns as $\boldsymbol{X}$ can be obtained.

Step 6: Replace the MV in $\boldsymbol{X}$ with the plausible values in the corresponding positions in $\boldsymbol{X}^{com}$ and use the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ in Step 1 to restore $\boldsymbol{X}$ to its original scale; hence, the final imputing result $\boldsymbol{X}^{imp}$ can be obtained.

It should be pointed out that the process of merging the same feature subsets is omitted in Step 3 for ease of understanding. Due to the subscript transformation that is likely to be confused during feature subsets merging, we think that an additional description herein may help understand this process. In fact, the features in $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ may be exactly the same if the correlation between $\boldsymbol{x}_{i,t}$ and $\boldsymbol{x}_{j,t}$ ($i,j \in \Omega$ and $i \neq j$) is great, that is, $\boldsymbol{X}_i$ and $\boldsymbol{X}_j$ are just different in the order of rows, but the corresponding values are exactly the same. In order to shorten the running time of the algorithm, it is necessary to merge the same feature subsets. The above steps constitute the core procedures of the DSCVBPCA imputation algorithm proposed in this paper. In order to facilitate the subsequent elaboration and the experiment-based superiority verification for the method that considers the future information on current samples, we rename the DSCVBPCA imputation algorithm with (17) fixed by $l \in [-f,0]$ as distributed delay correlation-based VBPCA (DDCVBPCA) imputation algorithm. Since the difference between DSCVBPCA and DDCVBPCA lies in the value of l, the DDCVBPCA algorithm can be implemented by a simple modification of the steps related to the value of l in DSCVBPCA.

3.4. Missing Values Imputation of Online Samples.

The main characteristics of MVI for online samples are as follows: (1) The data of online samples are obtained one by one over time, and the future information on current samples cannot be obtained. (2) Only the MV of current samples need to be imputed, whereas the MV of past samples needn't be processed. (3) The imputing model of online samples needs to be updated adaptively to track the latest state of process, and in this way, the high imputation performance can be maintained. (4) The imputation time of online samples must be less than the sampling periods of process data. Considering the above issues, we integrate the moving window strategy with DDCVBPCA method and obtain a new imputing algorithm named MWDDCVBPCA to accomplish the MVI of online samples. The main steps are as follows:

Input: An online sample $\hat{\boldsymbol{x}}_t \in \mathrm{R}^{d\times 1}$ with MV, the window size L, the maximum time shift coefficient f, and the cutoff value α of CPC.

Step 1: If there are MV in the online sample $\hat{\boldsymbol{x}}_t \in \mathrm{R}^{d\times 1}$ at the current moment t, construct the online data matrix $\hat{\boldsymbol{X}} = [\hat{\boldsymbol{x}}_{t-L+1}, \cdots, \hat{\boldsymbol{x}}_{t-1}, \hat{\boldsymbol{x}}_t] \in \mathrm{R}^{d\times L}$ via the moving window strategy. Then, calculate the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ of $\hat{\boldsymbol{X}}$ using the observed data, and convert $\hat{\boldsymbol{X}}$ to a matrix with zero mean and unit variance. Construct the time shift duplicate matrix $\hat{\boldsymbol{X}}_a = [\hat{\boldsymbol{X}}_{t-f}^T, \cdots, \hat{\boldsymbol{X}}_{t-1}^T, \hat{\boldsymbol{X}}_t^T]^T \in \mathrm{R}^{d(f+1)\times(L-f)}$ for $\hat{\boldsymbol{X}}$.

Step 2: Traverse all elements of $\hat{\boldsymbol{x}}_t$ and record the row coordinates where MV exist to the index set Ω.

Step 3: For each $i \in \Omega$, calculate its time shift correlation coefficient $\boldsymbol{r}_i = [r_{i,1}, r_{i,2}, \cdots, r_{i,d(f+1)-1}]$ with $l \in [-f,0]$ by (17), and sort the elements of $\boldsymbol{r}_i$ in descending order. Based on (20)−(21), select the feature subset $\boldsymbol{b}_i$ with stronger correlation with $\hat{\boldsymbol{x}}_{i,t}$ to get the best subset $\hat{\boldsymbol{X}}_i = [\hat{\boldsymbol{x}}_{i,t}^T, \hat{\boldsymbol{X}}_{\boldsymbol{b}_i}^T]^T$ to impute $\hat{\boldsymbol{x}}_{i,t}$. $\hat{\boldsymbol{x}}_{i,t}$ denotes the i-th row of $\hat{\boldsymbol{X}}_t$. Note that $\hat{\boldsymbol{X}}_{\boldsymbol{b}_i}$ is formed by selecting the rows of $\hat{\boldsymbol{X}}_a$ that correspond to the row coordinates recorded in $\boldsymbol{b}_i$.

Step 4: For each $i \in \Omega$, use VBPCA to impute $\hat{\boldsymbol{X}}_i$, and then the complete observation vector $\hat{\boldsymbol{x}}_{i,t}^{com}$ corresponding to $\hat{\boldsymbol{x}}_{i,t}$ can be obtained. Afterward, replace the i-th row of $\hat{\boldsymbol{X}}_t$ with $\hat{\boldsymbol{x}}_{i,t}^{com}$ to get the matrix $\hat{\boldsymbol{X}}_t^{com}$.

Step 5: Replace the MV in $\hat{\boldsymbol{x}}_t$ with the plausible values in the corresponding positions of the last column of $\hat{\boldsymbol{X}}_t^{com}$, and use the mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}$ in Step 1 to restore $\hat{\boldsymbol{x}}_t$ to its original scale; hence, the final imputing result $\hat{\boldsymbol{x}}_t^{imp}$ can be obtained.

Similar to Section 3.3, the process of merging the same feature subsets is omitted in Step 3 for ease of understanding. Because the online sample is obtained one by one over time and the future information on current sample cannot be obtained, the MVI of online samples adopts the DDCVBPCA algorithm instead of the DSCVBPCA algorithm. With the help of the moving window strategy to construct online data set, MWDDCVBPCA can adapt to the change of process states to a certain extent. It is noteworthy that the imputation performance of MWDDCVBPCA is related to the window size L. Although the model with a small MW size (MWS) may capture the changes of process quickly, it does not contain enough information to impute the MV of the current sample. Conversely, if the selected MWS is longer than the suitable one, a clumsy and insensitive model containing useless data may be produced. So far, choosing an appropriate L is still a tricky problem, especially for online modeling. At present, the most commonly used methods for window size selection are trial and error and empirical knowledge.[34]

3.5. Analysis of Time Complexity.

Without loss of generality, given an incomplete data set $\boldsymbol{X} \in R^{d\times n}$ containing d variables and n samples, the time complexity of the VBPCA executed on the whole data is $\mathrm{O}(d^2n^2)$.[41] For the proposed DSCVBPCA algorithm, the time complexity of two parts, namely, shift correlation coefficient and distributed VBPCA, should be taken into consideration. For any two variables (rows) in $\boldsymbol{X}$ and the maximum shift correlation coefficient f, the complexity of cross-correlations is $\mathrm{O}(fn)$ and thus the complexity of d variables is $\mathrm{O}(d^2fn)$. Supposing that the number of feature subsets is B and the number of variables in each feature subset is d_b(generally, $d_b < d$), the time complexity of distributed VBPCA is $\mathrm{O}(Bd_b^2n^2)$. As a result, the total time complexity of DSCVBPCA is $\mathrm{O}(Bd_b^2n^2 + d^2fn)$.

Remark 2. For the fast industrial process that may include large data sets, implementing the distributed MVI scheme developed in this work on a parallel platform can better satisfy the real-time requirement. More specifically, considering the B number of feature subsets in the distributed modeling framework for PWIP, the major computing task can be divided and assigned to the B number of CPUs. Each CPU trains the VBPCA for each feature subset to save the computation time.[42] In this way, the total imputation time will be equal to the time required to impute the feature subset with the longest imputation time, without being affected by the number of feature subsets. The ability of the designed distributed MVI solution in dealing with high dimensional data is reflected in two aspects: (i) The devised algorithm can

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

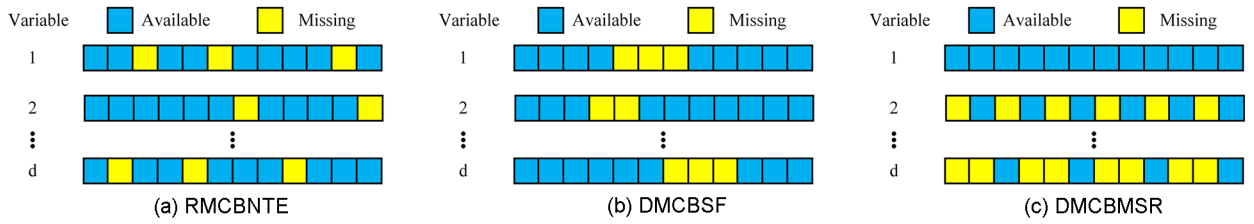


Figure 3. Common missingness mechanisms in industry process.

be implemented on the current mainstream parallel platforms; (ii) The MVI for training data sets is performed offline and has a moderate real-time requirement.

Remark 3. Before establishing the imputation model, we first select the feature subsets. When d is large, d_b is a relatively smaller value compared to d, and the lower data dimension d_b will reduce the demand for the amount of data n,[43] thereby reducing the imputation time of the submodel. From the perspective of practical applications, online imputation time can be further reduced from the following two aspects: (i) Executing the devised imputing algorithm on a parallel platform will greatly reduce the computation time. (ii) For the challenging situation where both data dimension and real-time requirement of the process are high, a base model with lower computational complexity can be adopted to replace VBPCA. This actually makes a proper trade-off between imputation accuracy and computational complexity, thereby expanding the applicable scope of online samples MVI.

4. SIMULATION STUDY

4.1. Setup and Design of Experiments. According to ref 44, there are three types of missingness mechanisms that can generate an incomplete data set, i.e., missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). As the research background of this work, the gold hydrometallurgy process (GHP) is one of the typical chemical industries, and unlike many other fields, its missingness mechanism is mainly MCAR or NMAR. Herein, NMAR is not considered because on the one hand, it is almost impossible to prove that the missingness mechanism is NMAR without knowing the missing elements, and on the other hand, NMAR is usually caused by data collection rules (e.g., elements above or under certain value are not registered).[45] Generally speaking, the types of MV in process industries mainly include the following: (1) Random missing caused by network transmission errors (RMCBNTE), etc. The positions of MV are random and irregular, as shown in Figure 3a. (2) Data missing caused by sensor failures (DMCBSF). These MV usually appear as small groups of sequential points lost at one time, but the groups are scattered randomly, as shown in Figure 3b. In the actual industry, workers will check and repair the equipment in time, so sensor failures are not regarded to be long-term or permanent. Based on this, the maximum consecutive missing elements in the subsequent experiment are set to 10. (3) Data missing caused by multiple sampling rates (DMCBMSR). Multiple sampling rates in the process are usually caused by the long-time offline test of some quality-related variables, as shown in Figure 3c. Although the above three types of MV are classified as MCAR, their modes are still different.[46] Since the soft measurement technology[47,48] is more suitable for the MVI of DMCBMSR, we only deal with RMCBNTE and DMCBSF in this section.

In practical applications, the MV in the data set are unknown, so it is difficult to evaluate the performances of different imputing algorithms. In order to generate the missing data set conforming to RMCBNTE or DMCBSF from the complete data set, the existing achievements usually select the positions of MV and the number of consecutive missing elements randomly. Although this approach is feasible, it lacks a corresponding inspection mechanism. Fortunately, ref 46 provides a tool for generating missing data set of RMCBNTE and DMCBSF, which is adopted herein. During the experiments, MV are artificially added through the corresponding missing mechanism. The closer the imputed values to the true values of MV, the better the imputation effect. Therefore, the widely used mean absolute error (MAE) and root-mean-square error (RMSE)[13,33] are exploited to evaluate the performances of imputing algorithms. Formulas of MAE and RMSE are given as follows:

$$MAE = \frac{1}{|\Psi|}\sum_{(i,j)\in\Psi}|x_{ij}^{imp} - x_{ij}^{true}| \tag{22}$$

$$RMSE = \sqrt{\frac{1}{|\Psi|}\sum_{(i,j)\in\Psi}(x_{ij}^{imp} - x_{ij}^{true})^2} \tag{23}$$

where Ψ denotes the set of MV coordinates in the data set, $|\Psi|$ denotes the number of elements in Ψ, and x_{ij}^{imp} and x_{ij}^{true} denote the imputed value and the true value, respectively. Obviously, smaller MAE and RMSE indicate better imputation performance. In general, a higher missing rate means that more information is lost, making it more difficult to accomplish MVI. For the purpose of ensuring production safety and economic benefits of industrial processes, workers are required to maintain equipment as soon as possible, so the missing rate of the data set will not be too high. Three representative cases where the missing rates are respectively 10%, 20%, and 30% are tested in this section. The formula of missing rate is as follows:[1]

$$\Upsilon = \frac{|\Psi|}{nd} \times 100\% \tag{24}$$

In order to verify the effectiveness and superiority of the proposed solution, the designed algorithm will be compared with the following six representative imputing algorithms:

Mean. MV are imputed via calculating the mean of nonmissing values of a variable.

kNN.[19] The average of k nearest-neighbors obtained by Euclidean distance is used for MVI. It is verified through a large number of experiments that k = 10 is a referential choice.[11]

LRMC.[20] LRMC is implemented by finding or approximating a low-rank matrix based on observable entries of the incomplete matrix. Rank minimization can be relaxed as the

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

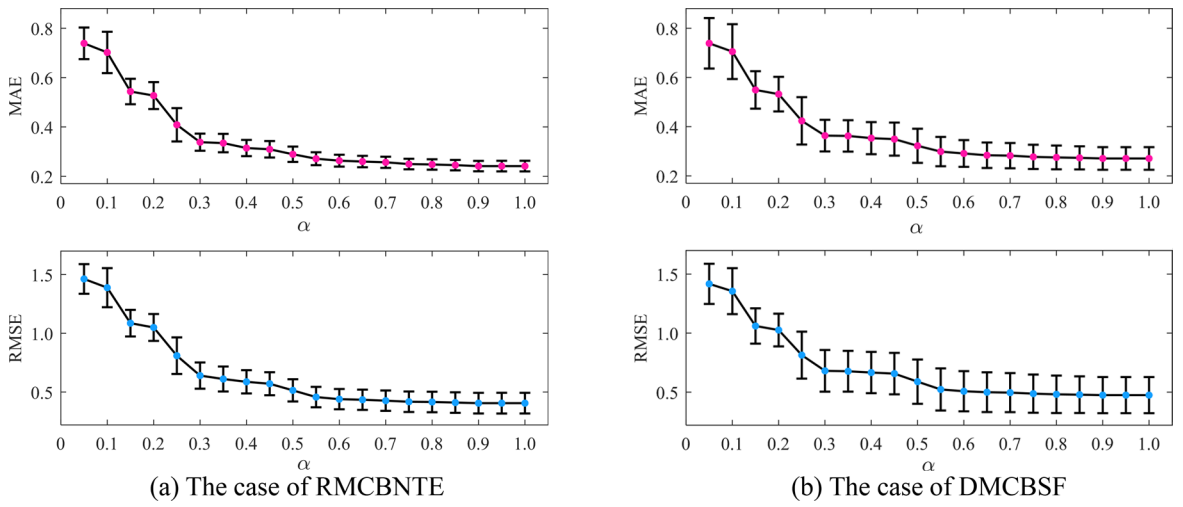


Figure 4. Effect of different values of α on the mean (circles) and standard deviation (bars) of imputation error.

problem of nuclear norm optimization, which can be efficiently solved by an algorithm called fixed point continuation with approximate singular value decomposition. Parameters are set according to "hard" problems.

EM.[21] EM imputation relies on the iterated analysis of linear regression of the variables with MV over those with available values. Regression coefficients are estimated by ridge regression and default values are used for parameters, e.g., stagnation tolerance = 0.005, maximum number of iterations = 30.

PPCA and VBPCA.[22] PPCA can be regarded as a maximum-likelihood reformulation of PCA. VBPCA introduces prior distributions over model parameters to handle the overfitting problem. Hyperparameters can be updated by variational Bayesian learning. In the experiments, the number of PCs is $c = d-1$ and the maximum number of iterations is 30.

DDCVBPCA and DSCVBPCA Proposed in This Paper. The cutoff value α is mainly determined by means of experimenting. In actual production processes, the MV are unknown. To find out the best value of α, a certain amount of MV can be artificially added to the observed data in the incomplete data set $\boldsymbol{X}$. If the missing rate of $\boldsymbol{X}$ is low (less than 10%), artificially add a larger missing rate (e.g., 20%) to the observed data to make the selected α more general. In contrast, artificially add a smaller missing rate (e.g., 10%) to ensure that the time shift correlation can be effectively calculated. With the artificially added MV, use the proposed methods to impute the MV, and adopt MAE and RMSE to evaluate the imputation performance. Select the corresponding α when both MAE and RMSE take their minimum values. To balance the imputation accuracy and the computational complexity when the data dimension is high, select the corresponding α when both MAE and RMSE take smaller values. It can be concluded from experiments that $\alpha \in [0.7,0.9]$ makes a satisfactory imputation performance. The remaining parameters are set as follows: the maximum time lags $f = 1$, the number of PCs = (the number of variables in each feature subset) − 1, and the maximum number of iterations is 30.

For MVI of online samples, the MW strategy is exploited to update the online data set via adding the newly acquired sample and removing the oldest sample, so that the latest process information is introduced into the imputation model without increasing computation burden. An appropriate selection of MWS is important because the MWS has an impact on the sensitivity of MVI. Given this, the MWS herein is determined by trial and error. That is, select the MWS when both MAE and RMSE take smaller values through comparing imputation performances under MWs of different sizes. Due to the length limitation of the paper, the detailed selection process of MW is omitted. All algorithms are implemented by MATLAB 2018b on a Windows computer with a Core i7 3.6 GHz processor and 8 GB RAM.

4.2. Numerical Example. The numerical simulation is performed on a modified system based on the following:[49]

$$\boldsymbol{z}_1(t) = \begin{bmatrix} 0.15 & -0.12 \\ -0.12 & 0.19 \end{bmatrix}\boldsymbol{z}_1(t-1) + \begin{bmatrix} -0.56 & 0.51 \\ 0.59 & -0.53 \end{bmatrix}\boldsymbol{u}_1(t-1) \tag{25}$$

$$\boldsymbol{z}_2(t) = \begin{bmatrix} 0.85 & 0.33 \\ 0.25 & 0.79 \end{bmatrix}\boldsymbol{u}_2(t) \tag{26}$$

$$\boldsymbol{u}_1(t) = \begin{bmatrix} -0.39 & 0.48 \\ 0.32 & -0.44 \end{bmatrix}\boldsymbol{u}_1(t-1) + \begin{bmatrix} 0.61 & -0.22 \\ -0.55 & 0.15 \end{bmatrix}\boldsymbol{w}_1(t-1) \tag{27}$$

$$\boldsymbol{u}_2(t) = \begin{bmatrix} 0.35 & 0.49 \\ 0.42 & 0.59 \end{bmatrix}\boldsymbol{w}_2(t) \tag{28}$$

$$\boldsymbol{y}(t) = \begin{bmatrix} \boldsymbol{z}_1(t) \\ \boldsymbol{z}_2(t) \end{bmatrix} + \boldsymbol{v}(t) \tag{29}$$

where $\boldsymbol{w}$ is a random variable with the mean being 1 and variance being 3, and $\boldsymbol{v}$ is the process noise with the mean being 0 and variance being 0.1. The vector $\boldsymbol{x} = [\boldsymbol{y}^T, \boldsymbol{u}_1^T, \boldsymbol{u}_2^T]^T \in R^{8\times 1}$ of measured variables is used for evaluating the performance of imputing algorithms. Obviously, the system has eight measured variables and consists of two independent subsystems. Each subsystem can be deemed as a subprocess of the plant-wide process. Although the numerical example is not an actual large-scale system with physical meanings, it can verify the imputation performance intuitively.

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

Table 1. Performances of Different Imputing Algorithms for RMCBNTE on the Training Dataset

	10%			20%			30%		
	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)
Mean	2.030±0.101	2.617±0.118	7.4×10^{-4}	2.035±0.065	2.625±0.078	7.7×10^{-4}	2.038±0.054	2.630±0.061	7.9×10^{-4}
kNN	0.643±0.059	0.956±0.116	0.449	0.830±0.051	1.267±0.095	0.811	1.050±0.056	1.580±0.093	1.086
LRMC	0.465±0.157	0.716±0.230	0.827	0.572±0.149	0.909±0.195	0.817	0.669±0.138	1.071±0.181	0.813
EM	0.674±0.064	0.919±0.088	5.394	0.418±0.041	0.736±0.097	7.900	0.643±0.041	1.007±0.085	8.976
PPCA	0.280±0.038	0.499±0.120	0.318	0.356±0.036	0.695±0.104	0.310	0.439±0.040	0.866±0.100	0.305
VBPCA	0.274±0.039	0.491±0.123	0.392	0.351±0.036	0.688±0.103	0.378	0.437±0.041	0.860±0.101	0.367
DDCVBPCA	0.226±0.025	0.378±0.092	0.907	0.270±0.028	0.511±0.095	0.993	0.326±0.032	0.629±0.088	1.033
DSCVBPCA	**0.202±0.012**	**0.274±0.033**	1.437	**0.216±0.014**	**0.333±0.076**	1.264	**0.242±0.022**	**0.403±0.085**	1.254

Table 2. Performances of Different Imputing Algorithms for DMCBSF on the Training Dataset

	10%			20%			30%		
	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)
Mean	2.034±0.176	2.612±0.221	7.5×10^{-4}	2.040±0.103	2.621±0.124	8.4×10^{-4}	2.035±0.083	2.621±0.110	7.4×10^{-4}
kNN	0.645±0.081	0.953±0.127	0.436	0.804±0.075	1.215±0.130	0.798	1.038±0.098	1.564±0.173	1.034
LRMC	0.489±0.187	0.744±0.290	0.837	0.521±0.165	0.823±0.226	0.814	0.647±0.171	1.040±0.246	0.800
EM	0.644±0.083	0.883±0.106	5.269	0.417±0.083	0.695±0.144	8.000	0.601±0.099	0.960±0.145	8.916
PPCA	0.286±0.061	0.499±0.151	0.318	0.334±0.056	0.630±0.140	0.318	0.433±0.069	0.846±0.149	0.301
VBPCA	0.278±0.059	0.487±0.154	0.394	0.330±0.055	0.622±0.140	0.386	0.430±0.068	0.840±0.148	0.365
DDCVBPCA	0.232±0.041	0.385±0.125	0.775	0.267±0.048	0.478±0.138	0.785	0.339±0.057	0.660±0.153	0.847
DSCVBPCA	**0.204±0.020**	**0.276±0.036**	1.339	**0.220±0.025**	**0.324±0.084**	1.243	**0.256±0.039**	**0.443±0.145**	1.295

The training data set containing 300 samples generated by the above system is used to verify the imputation performance of DSCVBPCA for the incomplete data set. Since the MV in actual productions are unknown, the following steps present a detailed experimental process. First, artificially add 10% MV to the training data set through the corresponding missingness mechanism, and use this incomplete data set to simulate the one obtained from the actual application; that is, assume that the MV in this data set are unknown. Second, again artificially add 20% MV to the observations of the incomplete data set. Note that the true values of these MV are known. Finally, impute the MV via the proposed method and use MAE and RMSE to evaluate the imputation effect of DSCVBPCA for the MV added at the second time. To make the experimental results more general, for different values of α, the procedure of artificially adding MV and imputing MV is repeated 50 times. Figure 4 shows the effect of different values of α on the mean and standard deviation of imputation error. We can see from Figure 4 that MAE and RMSE show a downward trend when $\alpha \in (0,0.9]$ and achieve the minimum value at $\alpha = 0.9$, but when $\alpha \in (0.9,1]$, there is no significant change in MAE and RMSE. This is mainly because the correlations of features of the subprocesses in the numerical example are relatively large, and when $\alpha \geq 0.9$, the feature subsets are fixed. Therefore, α is taken as 0.9 in the subsequent experiments.

To make the experimental results more general, in the subsequent experiments, we will generate 100 missing data sets based on the training set through corresponding missingness mechanisms, then test different imputing algorithms, and count the mean and standard deviation of MAE and RMSE, respectively. The imputation performances (including imputation accuracy and imputation time) for the cases of RMCBNTE and DMCBSF are listed in Table 1 and Table 2, respectively, where the best results with highest imputation accuracy are marked in bold. Table 1 shows that, on the whole, as the missing rate increases, the imputation effect becomes worse, except for mean and EM algorithms. Next, this phenomenon is analyzed in detail. For the mean algorithm, when the missing rate is low and is not enough to affect the estimated value of mean, the imputation result will not make a significant difference. In addition, the utilization of the mean neglects the variance of data or correlations between variables. As a consequence, the imputation effect of mean algorithm is unsatisfactory. As for the EM algorithm, there exists approximate collinearity among measured variables (refer to $\boldsymbol{u}_2$), which is quite unfavorable for parameter estimation of EM imputing model, and even leads to the nonconvergence of the algorithm. Therefore, the imputation effect with a missing rate of 20% seems to be better than the case 10%. Nevertheless, EM is still a strong competitor of PPCA and VBPCA. The performance of kNN is relatively good for a stationary data set with a low missing rate; however, as the missing rate increases, the performance degrades rapidly. In this case, an algorithm that takes into account the correlations between variables will achieve a relatively stable imputation effect. Although DDCVBPCA only uses the current and past information, its imputation performance has been significantly improved compared with VBPCA. Since DSCVBPCA further utilizes the future information, its performance is improved with a certain degree compared to DDCVBPCA. When Table 1 and Table 2 are compared, it can be observed that the imputation effect of DDCVBPCA and DSCVBPCA for randomly MV is better than that of consecutive MV. This is mainly because the consecutive missing will destroy the temporal correlations of samples, and meanwhile make some samples fail to use the autocorrelation information. Obviously, the average imputation time (i.e., the total imputation time of 100 incomplete data sets/100) of the EM algorithm is the longest because of two reasons: on the one hand, the computational complexity of EM is relatively high; on the other hand, there exists approximate collinearity among some variables, leading to a longer parameter-estimation time of EM imputing model. In addition, missing rates have greater impact on the average imputation time of EM and kNN, but have less impact on the average imputation time of other algorithms. This is related to the imputation principles of different imputing algorithms. For example, the more incomplete samples caused by a larger missing rate will increase the number of times that kNN needs

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

Table 3. Performances of Different Imputing Algorithms for RMCBNTE on the Online Samples

	10%			20%			30%		
	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)
MWMean	1.934±0.105	2.412±0.123	3.3×10^{-4}	1.956±0.061	2.441±0.073	3.3×10^{-4}	1.948±0.052	2.433±0.070	3.1×10^{-4}
MWkNN	0.777±0.054	1.135±0.108	1.1×10^{-3}	0.885±0.050	1.292±0.089	1.4×10^{-3}	1.006±0.051	1.466±0.094	1.5×10^{-3}
MWLRMC	0.484±0.049	0.746±0.105	0.576	0.584±0.044	0.901±0.102	0.583	0.690±0.055	1.079±0.111	0.579
MWEM	0.667±0.057	0.899±0.088	1.319	0.509±0.042	0.778±0.089	1.933	0.612±0.046	0.945±0.090	2.180
MWPPCA	0.289±0.041	0.488±0.115	0.142	0.360±0.036	0.658±0.102	0.140	0.436±0.045	0.823±0.110	0.140
MWVBPCA	0.285±0.037	0.481±0.111	0.166	0.357±0.034	0.650±0.102	0.165	0.440±0.044	0.816±0.108	0.161
MWDDCVBPCA	**0.240±0.029**	**0.391±0.106**	0.171	**0.279±0.031**	**0.503±0.100**	0.231	**0.343±0.034**	**0.654±0.102**	0.339

Table 4. Performances of Different Imputing Algorithms for DMCBSF on the Online Samples

	10%			20%			30%		
	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)
MWMean	1.937±0.142	2.408±0.181	3.5×10^{-4}	1.951±0.108	2.429±0.141	3.3×10^{-4}	1.943±0.078	2.422±0.094	3.1×10^{-4}
MWkNN	0.786±0.094	1.122±0.167	1.2×10^{-3}	0.892±0.083	1.298±0.148	1.3×10^{-3}	0.995±0.067	1.447±0.130	1.5×10^{-3}
MWLRMC	0.484±0.065	0.733±0.127	0.584	0.583±0.075	0.909±0.146	0.581	0.692±0.075	1.063±0.137	0.572
MWEM	0.593±0.072	0.816±0.105	1.306	0.551±0.069	0.823±0.130	2.003	0.585±0.063	0.908±0.125	2.152
MWPPCA	0.295±0.068	0.484±0.160	0.143	0.365±0.073	0.662±0.168	0.140	0.437±0.069	0.807±0.151	0.119
MWVBPCA	0.291±0.062	0.471±0.158	0.167	0.363±0.071	0.650±0.168	0.163	0.438±0.068	0.803±0.145	0.138
MWDDCVBPCA	**0.239±0.050**	**0.366±0.145**	0.180	**0.295±0.065**	**0.530±0.159**	0.233	**0.352±0.060**	**0.651±0.138**	0.276

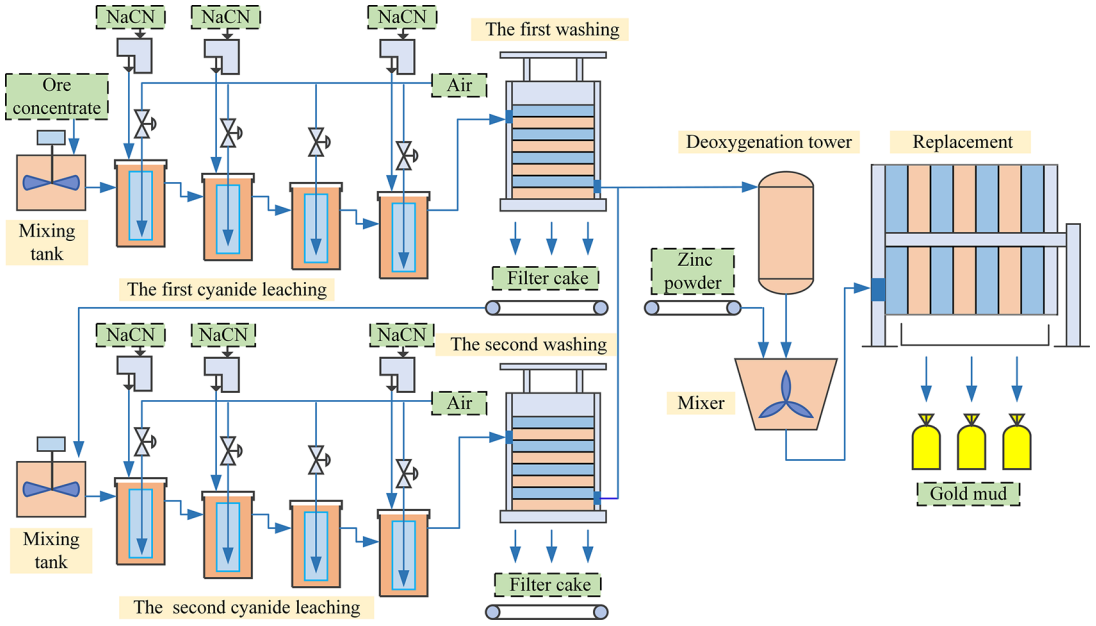


Figure 5. Flowchart of GHP.

to search the entire data set, and thus the imputation time will be longer. In contrast, PPCA and VBPCA use the data reconstituted by projection vectors and score vectors to replace the missing values, and even if there is only one missing value in the data set, they still need to reconstitute the entire data set, so the missing rate has less impact on them.

The MVI of online samples is necessary for the industrial field because most of the intelligent decision support systems demand a complete training data set and online samples. Generate 300 samples by the system, and from the 101st sample, all elements of the coefficient matrix of $\boldsymbol{u}_1(t-1)$ in (27) are subtracted by 0.01 at each sampling moment to simulate the parameter drift of some variables. By means of trial and error, choose the MWS as 100 and, from the 101st sample, add MV via the corresponding missingness mechanisms. The procedure of artificially adding MV and imputing MV is repeated 100 times. The imputation performances for the cases of RMCBNTE and DMCBSF are listed in Table 3 and Table 4, respectively, where MWDDCVBPCA exhibits the best imputation effect. As we all know, the imputation time is very important for online samples, and a too long imputation time cannot meet the actual requirements. Despite the best imputation accuracy of MWDDCVBPCA, its average imputation time (i.e., the total imputation time of online samples with MV/the total number of online samples with MV) is longer than that of MWVBPCA, which is reasonable according to the analysis of time complexity in Section 3.5. It is worth mentioning that we have merged the same feature subsets, which saves the computation time of repeated subsets, otherwise, the imputation time is likely to be longer. Besides, the online imputation time of MWDDCVBPCA increases as the missing rate increases. This is mainly because the more serious the missing rate, the more missing values exist in online samples (also indicating the greater number of feature subsets),

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

resulting in the longer imputation time. Fortunately, MWDDCVBPCA is qualified to be run in parallel. If the algorithm is executed on a distributed running platform, the total online imputation time will be equal to the time required to impute the feature subset with the longest imputation time. In this way, the imputation time can be significantly reduced without being affected by the number of missing values present in online samples.

4.3. Application to the Gold Hydrometallurgy Process. The hydrometallurgy process is an efficient technology of extracting gold from low-grade ores. In this subsection, the feasibility and effectiveness of the proposed MVI solution are verified on the semi-physical simulation platform of the gold hydrometallurgy process (GHP) that provides the foundation for development and debugging of process monitoring, operating performance assessment, and MVI.[50,51] The typical GHP is composed of five fundamental units: the first cyanide leaching, the first washing, the second cyanide leaching, the second washing, and replacement. The flowchart of GHP is shown in Figure 5. The process of the first cyanide leaching is to mix the gold-containing ore particles with particle size being about 400 mesh with the barren liquor, and blend the resulting mixture to be the ore pulp with a concentration of 25% to 35%. Afterward, add sodium cyanide (NaCN) to the resulting ore pulp and fill it with air. In this way, the gold in ore can fully react with NaCN and finally appears as $[Au(CN)_2]^-$ in the liquid phase. The chemical reaction is given by (30). To avoid the highly toxic gas HCN generated by the hydrolysis of NaCN, add calcium hydroxide ($Ca(OH)_2$) to adjust the pH of ore pulp to about 11. After treatment of the first cyanide leaching, leached ore pulp is transferred to an automatic vertical filter press to accomplish the separation of pregnant solution and filter cake, so as to separate the leached $[Au(CN)_2]^-$ timely. For the purpose of minimizing the amount of $[Au(CN)_2]^-$ that is attached to filter cake, it is necessary to repeatedly wash filter cake with barren liquor, and this process is called the first washing. The objective of the second cyanide leaching and second washing that respectively correspond to the first cyanide leaching and first washing is to extract as much gold as possible. Replacement is conducted when the separated pregnant solution flows into the frame filter press after deoxygenation processing. In the end, add zinc powder to the pregnant solution to extract the solid gold. The chemical reaction is given by (31).

$$4Au + 8NaCN + 2H_2O + O_2 = 4NaAu(CN)_2 + 4NaOH \quad (30)$$

$$2NaAu(CN)_2 + Zn = 2Au + Na_2Zn(CN)_4 \quad (31)$$

Based on the GHP simulation platform, the process variables used for MVI are listed in Table 5. The training data set containing 500 samples generated by GHP is used to verify the imputation performance of DSCVBPCA for an incomplete data set. The determination of α is the same as that in Section 4.2. Figure 6 shows the effect of different values of α on the mean and standard deviation of imputation error. We can see from Figure 6 that MAE and RMSE show a downward trend when $\alpha \in (0,0.7)$, there is no significant change in MAE and RMSE when $\alpha \in [0.7,0.9]$, and MAE and RMSE gradually increase when $\alpha \in (0.9,1]$. This is mainly because the correlations of some variables are not uniform; that is, the

Table 5. Variables Used for MVI of GHP

No.	Subprocess	State
1	The first cyanide leaching	concentration of ore pulp
2		inlet pulp flow
3		NaCN flow 1
4		NaCN flow 2
5		NaCN flow 4
6		air flow
7		concentration of dissolved oxygen
8		concentration of CN^- 1
9		concentration of CN^- 2
10		concentration of CN^- 4
11	The first washing	feed pressure of vertical filter press
12		extrusion pressure of vertical filter press
13		hydraulic pressure of vertical filter press
14	The second cyanide leaching	inlet pulp flow
15		NaCN flow 1
16		NaCN flow 2
17		NaCN flow 4
18		air flow
19		concentration of dissolved oxygen
20		concentration of CN^- 1
21		concentration of CN^- 2
22		concentration of CN^- 4
23	The second washing	feed pressure of vertical filter press
24		extrusion pressure of vertical filter press
25		hydraulic pressure of vertical filter press
26	Replacement	vacuum degree of deaeration tower
27		$[Au(CN)_2]^-$ concentration of precious liquid
28		$[Au(CN)_2]^-$ concentration of barren liquor
29		adding amount of zinc powder
30		hydraulic pressure of frame filter press

variables with MV have a greater correlation with some features, but have a smaller correlation with other features. When there are already large-correlation features in the subset, further addition of small-correlation features will reduce the accuracy of the algorithm and increase the computational burden. Due to the relatively large number of variables in GHP, after comprehensively weighing the imputation accuracy and computational complexity, α is taken as 0.7 in the subsequent experiments.

To make the experimental results more general, we will generate 100 missing data sets based on the training set through corresponding missingness mechanisms, then test different imputing algorithms, and count the mean and standard deviation of MAE and RMSE, respectively. The imputation performances for the cases of RMCBNTE and DMCBSF are listed in Table 6 and Table 7, respectively, where the proposed imputing algorithm achieves the best imputation accuracy. It can be found that the imputation performance of DDCVBPCA and DSCVBPCA for DMCBSF degrades more than that for RMCBNTE, which reflects the fact that the autocorrelations of variables cannot be ignored in the actual industrial process. Compared to VBPCA, the imputation accuracy of the designed DDCVBPCA and DSCVBPCA is greatly improved, but the computation time is relatively longer.

Generate 500 online samples by GHP, and from the 201st sample, add the random disturbance $N(0, 0.02^2)$ to the concentration of ore pulp (variable 1). Under the adjustment

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

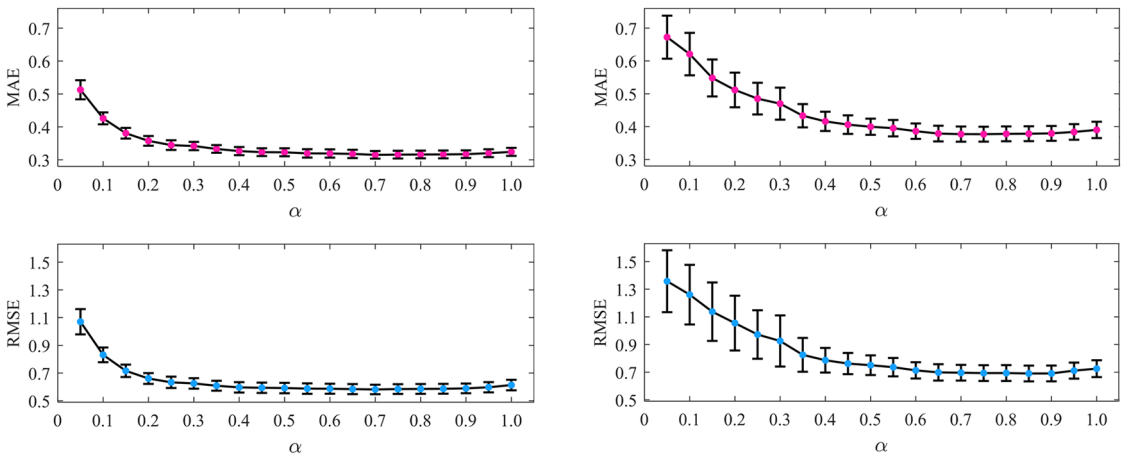


Figure 6. Effect of different values of α on the mean (circles) and standard deviation (bars) of imputation error for GHP.

Table 6. Performances of Different Imputing Algorithms for RMCBNTE of GHP on the Training Dataset

	10%			20%			30%		
	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)
Mean	1.735±0.062	2.922±0.110	2.2×10^{-3}	1.733±0.041	2.928±0.073	2.1×10^{-3}	1.729±0.035	2.918±0.058	2.5×10^{-3}
kNN	0.905±0.033	1.384±0.056	5.154	0.963±0.024	1.499±0.044	9.638	1.034±0.025	1.633±0.047	12.582
LRMC	0.764±0.076	1.201±0.141	4.728	0.855±0.062	1.357±0.113	4.499	0.967±0.075	1.539±0.128	4.143
EM	0.432±0.025	0.732±0.053	10.448	0.425±0.016	0.766±0.042	6.210	0.530±0.019	0.925±0.046	8.293
PPCA	0.329±0.016	0.580±0.062	1.847	0.407±0.016	0.751±0.045	1.740	0.490±0.020	0.912±0.051	1.681
VBPCA	0.320±0.018	0.576±0.060	3.010	0.400±0.016	0.746±0.043	2.794	0.488±0.019	0.910±0.049	2.610
DDCVBPCA	0.244±0.013	0.418±0.054	18.993	0.297±0.011	0.537±0.035	19.375	0.359±0.012	0.655±0.035	20.155
DSCVBPCA	**0.215±0.010**	**0.372±0.038**	28.279	**0.266±0.010**	**0.485±0.032**	27.673	**0.323±0.010**	**0.586±0.029**	27.940

Table 7. Performances of Different Imputing Algorithms for DMCBSF of GHP on the Training Dataset

	10%			20%			30%		
	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)
Mean	1.741±0.158	2.929±0.285	2.3×10^{-3}	1.749±0.111	2.955±0.199	2.3×10^{-3}	1.745±0.084	2.944±0.147	2.4×10^{-3}
kNN	0.948±0.068	1.461±0.111	5.101	1.012±0.055	1.585±0.095	9.499	1.076±0.050	1.715±0.100	12.462
LRMC	0.773±0.088	1.216±0.167	4.600	0.904±0.082	1.438±0.152	4.436	1.013±0.080	1.622±0.141	4.128
EM	0.438±0.039	0.736±0.087	11.709	0.434±0.031	0.785±0.076	7.122	0.544±0.033	0.942±0.076	7.896
PPCA	0.329±0.031	0.586±0.091	1.812	0.412±0.032	0.760±0.086	1.748	0.495±0.031	0.930±0.087	1.627
VBPCA	0.327±0.033	0.577±0.090	2.988	0.407±0.030	0.758±0.084	2.857	0.491±0.033	0.923±0.082	2.607
DDCVBPCA	0.260±0.028	0.448±0.088	17.212	0.332±0.026	0.612±0.076	18.656	0.410±0.029	0.768±0.075	18.597
DSCVBPCA	**0.237±0.023**	**0.412±0.080**	27.941	**0.302±0.024**	**0.558±0.068**	27.495	**0.374±0.024**	**0.697±0.062**	26.548

Table 8. Performances of Different Imputing Algorithms for RMCBNTE of GHP on the Online Samples

	10%			20%			30%		
	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)
MWMean	9.945±0.632	23.426±1.431	6.3×10^{-4}	9.739±0.450	22.996±1.052	6.6×10^{-4}	9.997±0.262	23.439±0.537	6.7×10^{-4}
MWkNN	4.272±0.342	9.461±0.758	3.3×10^{-3}	4.544±0.175	10.226±0.479	5.9×10^{-3}	5.096±0.243	11.503±0.685	8.2×10^{-3}
MWLRMC	2.944±0.268	6.795±0.804	2.026	3.277±0.160	7.474±0.483	1.998	3.944±0.181	9.141±0.610	1.971
MWEM	1.874±0.133	4.429±0.381	2.061	1.581±0.069	3.652±0.251	1.729	1.598±0.085	3.816±0.353	2.108
MWPPCA	0.896±0.061	2.055±0.123	0.619	1.146±0.061	2.553±0.185	0.604	1.413±0.070	3.042±0.239	0.586
MWVBPCA	0.911±0.049	2.121±0.112	0.985	1.133±0.059	2.446±0.191	0.943	1.381±0.073	2.958±0.249	0.907
MWDDCVBPCA	**0.422±0.027**	**0.760±0.084**	0.709	**0.497±0.028**	**0.918±0.065**	1.372	**0.615±0.044**	**1.191±0.117**	2.171

Table 9. Performances of Different Imputing Algorithms for DMCBSF of GHP on the Online Samples

	10%			20%			30%		
	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)	MAE	RMSE	Time(s)
MWMean	9.635±1.837	22.412±3.640	6.1×10^{-4}	10.571±1.288	24.623±2.751	6.3×10^{-4}	10.714±1.048	24.915±2.041	6.5×10^{-4}
MWkNN	5.415±1.503	12.206±3.212	3.2×10^{-3}	5.985±0.783	13.692±1.796	5.6×10^{-3}	6.386±0.642	14.573±1.614	8.1×10^{-3}
MWLRMC	3.584±0.785	8.251±2.106	2.017	4.495±0.586	10.647±1.643	1.981	5.155±0.626	12.241±1.769	1.973
MWEM	2.297±0.527	5.396±1.299	3.068	2.185±0.333	5.092±0.917	2.176	2.227±0.373	5.203±1.161	2.540
MWPPCA	1.373±0.283	2.685±0.813	0.614	1.515±0.269	3.499±0.758	0.603	1.825±0.317	4.356±1.057	0.587
MWVBPCA	1.381±0.295	2.697±0.827	0.988	1.534±0.275	3.526±0.771	0.943	1.774±0.291	4.302±1.015	0.905
MWDDCVBPCA	**0.622±0.111**	**1.220±0.404**	0.707	**0.888±0.096**	**2.053±0.339**	1.347	**1.209±0.185**	**3.026±0.766**	2.167

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593

of the closed-loop controller, small-amplitude fluctuations appear in the process. Obviously, the current process exhibits strong dynamic characteristics. By means of trial and error, choose the MWS as 150, and from the 151st sample, add MV via the corresponding missingness mechanisms. The procedure of artificially adding MV and imputing MV is repeated 100 times. The imputation performances for the cases of RMCBNTE and DMCBSF are listed in Table 8 and Table 9, respectively, where MWDDCVBPCA exhibits the best imputation effect. The average imputation time of each of the seven online imputing algorithms for each sample is acceptable since the sampling period of GHP is 60 s. When the missing rate is 10%, compared to MWVBPCA, the developed MWDDCVBPCA not only has a shorter imputation time but also has higher imputation accuracy. The reasons are twofold: on the one hand, a smaller missing rate indicates fewer missing values in online samples (also indicates fewer feature subsets), and on the other hand, MWDDCVBPCA undergoes the procedure of feature selection. This experimental result also demonstrates that a suitable feature selection algorithm can not only reduce the data dimension but also improve the model performance.

5. CONCLUSIONS

In this work, an improved feature subset selection algorithm based on the time shift correlation and a newly defined selection criterion is developed. With the combination of this feature subset selection algorithm and VBPCA method, a missing values imputation technique suitable for plant-wide industrial processes is proposed. The designed strategy comprehensively considers the static and dynamic characteristics among process variables and makes full use of the information related to missing values. From the perspective of practical applications, online missing values imputation is also explored. By setting parameters of the presented algorithm elaborately and using the moving window method, the missing values of both training data set and online samples can be imputed with high precision. It is validated by the experiments on numerical example and gold hydrometallurgy that the proposed solution is feasible and effective. However, there are still two limitations that need further consideration: (1) Under the premise of not significantly increasing the computational complexity, it is expected that the correlations among variables will be expanded from linear relationships to the nonlinear relationships that are more suitable for actual industrial processes. (2) It is difficult for the moving window to handle the strong nonlinear process with abrupt changes, and solutions for this situation are desired. In future work, we will implement the developed solution on a parallel platform to further verify the effectiveness of the MVI scheme not only in improving the imputation accuracy but also in reducing the imputation time.

AUTHOR INFORMATION

Corresponding Author

Yuqing Chang − *College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China;* orcid.org/0000-0003-2010-0896; Email: changyuqing@ise.neu.edu.cn

Authors

Linsheng Zhong − *College of Information Science and Engineering, Northeastern University, Shenyang 110819, China;* orcid.org/0000-0003-4445-1594

Fuli Wang − *College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; State Key Laboratory of Synthetical Automation for Process Industries, Northeastern University, Shenyang 110819, China*

Shihong Gao − *College of Information Science and Engineering, Northeastern University, Shenyang 110819, China; School of Automation and Software Engineering, Shanxi University, Taiyuan 030006, China;* orcid.org/0000-0002-2367-8258

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.iecr.1c03860

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (Nos. 61873053, 61973057, and 61533007), the National Key Research and Development Program of China (2019YFE0105000), and the Foundation for Innovative Research Groups of the National Natural Science Foundation of China (61621004).

REFERENCES

(1) Pan, J.; Li, C.; Tang, Y.; Li, W.; Li, X. Energy Consumption Prediction of a CNC Machining Process With Incomplete Data. *IEEE/CAA J. Autom. Sin.* **2021**, *8* (5), 987−1000.

(2) Tao, Y.; Shi, H.; Song, B.; Tan, S. Distributed Supervised Fault Detection and Diagnosis for a Non-Gaussian Process. *Ind. Eng. Chem. Res.* **2019**, *58* (16), 6592−6603.

(3) Zhao, C.; Chen, J.; Jing, H. Condition-Driven Data Analytics and Monitoring for Wide-Range Nonstationary and Transient Continuous Processes. *ieee Trans. Autom. Sci. Eng.* **2021**, *18* (4), 1563−1574.

(4) Du, J.; Hu, M.; Zhang, W. Missing Data Problem in the Monitoring System: A Review. *IEEE Sens. J.* **2020**, *20* (23), 13984−13998.

(5) Zhu, J.; Ge, Z.; Song, Z.; Gao, F. Review and Big Data Perspectives on Robust Data Mining Approaches for Industrial Process Modeling with Outliers and Missing Data. *Annu. Rev. Control* **2018**, *46*, 107−133.

(6) Yuan, X.; Ge, Z.; Song, Z.; Wang, Y.; Yang, C.; Zhang, H. Soft Sensor Modeling of Nonlinear Industrial Processes Based on Weighted Probabilistic Projection Regression. *IEEE Trans. Instrum. Meas.* **2017**, *66* (4), 837−845.

(7) Strike, K.; El Emam, K. El; Madhavji, N. Software Cost Estimation with Incomplete Data. *IEEE Trans. Softw. Eng.* **2001**, *27* (10), 890−908.

(8) Luo, L.; Bao, S.; Peng, X. Robust Monitoring of Industrial Processes Using Process Data with Outliers and Missing Values. *Chemom. Intell. Lab. Syst.* **2019**, 192.

(9) Liu, Y.; Pan, Y.; Sun, Z.; Huang, D. Statistical Monitoring of Wastewater Treatment Plants Using Variational Bayesian PCA. *Ind. Eng. Chem. Res.* **2014**, *53* (8), 3272−3282.

(10) Zhu, J.; Ge, Z.; Song, Z. Robust Modeling of Mixture Probabilistic Principal Component Analysis and Process Monitoring Application. *AIChE J.* **2014**, *60* (6), 2143−2157.

(11) Sahri, Z.; Yusof, R.; Watada, J. FINNIM: Iterative Imputation of Missing Values in Dissolved Gas Analysis Dataset. *IEEE Trans. Ind. informatics* **2014**, *10* (4), 2093−2102.

(12) Wang, Z.; Wang, L.; Tan, Y.; Yuan, J. Fault Detection Based on Bayesian Network and Missing Data Imputation for Building Energy Systems. *Appl. Therm. Eng.* **2021**, *182*, 116051.

(13) He, D.; Wang, Z.; Yang, L.; Dai, W. Study on Missing Data Imputation and Modeling for the Leaching Process. *Chem. Eng. Res. Des.* **2017**, *124*, 1−19.
(14) Jiang, C.; Zhong, W.; Li, Z.; Peng, X.; Yang, M. Real-Time Semisupervised Predictive Modeling Strategy for Industrial Continuous Catalytic Reforming Process with Incomplete Data Using Slow Feature Analysis. *Ind. Eng. Chem. Res.* **2019**, *58* (37), 17406−17423.
(15) Li, D.; Li, L.; Li, X.; Ke, Z.; Hu, Q. Smoothed LSTM-AE: A Spatio-Temporal Deep Model for Multiple Time-Series Missing Imputation. *Neurocomputing* **2020**, *411*, 351−363.
(16) Choudhury, S. J.; Pal, N. R. Imputation of Missing Data with Neural Networks for Classification. *Knowl. based Syst.* **2019**, *182*, 104838.
(17) Zhang, S.; Jin, Z.; Zhu, X. Missing Data Imputation by Utilizing Information within Incomplete Instances. *J. Syst. Softw.* **2011**, *84* (3), 452−459.
(18) Huang, H.; Peng, X.; Jiang, C.; Li, Z.; Zhong, W. Variable-Scale Probabilistic Just-in-Time Learning for Soft Sensor Development with Missing Data. *Ind. Eng. Chem. Res.* **2020**, *59* (11), 5010−5021.
(19) Troyanskaya, O. G.; Cantor, M. N.; Sherlock, G.; Brown, P. O.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R. B. Missing Value Estimation Methods for DNA Microarrays. *Bioinformatics* **2001**, *17* (6), 520−525.
(20) Ma, S.; Goldfarb, D.; Chen, L. Fixed Point and Bregman Iterative Methods for Matrix Rank Minimization. *Math. Program.* **2011**, *128* (1), 321−353.
(21) Schneider, T. Analysis of Incomplete Climate Data: Estimation of Mean Values and Covariance Matrices and Imputation of Missing Values. *J. Clim.* **2001**, *14* (5), 853−871.
(22) Ilin, A.; Raiko, T. Practical Approaches to Principal Component Analysis in the Presence of Missing Values. *J. Mach. Learn. Res.* **2010**, *11* (66), 1957−2000.
(23) Blei, D. M.; Kucukelbir, A.; McAuliffe, J. D. Variational Inference: A Review for Statisticians. *J. Am. Stat. Assoc.* **2017**, *112* (518), 859−877.
(24) Chu, F.; Liang, T.; Chen, C. L. P.; Wang, X.; Ma, X. Weighted Broad Learning System and Its Application in Nonlinear Industrial Process Modeling. *IEEE Trans. neural networks* **2020**, *31* (8), 3017−3031.
(25) Ge, Z.; Chen, J. Plant-Wide Industrial Process Monitoring: A Distributed Modeling Framework. *IEEE Trans. Ind. informatics* **2016**, *12* (1), 310−321.
(26) Jiang, J.; Jiang, Q. Variational Bayesian Probabilistic Modeling Framework for Data-Driven Distributed Process Monitoring. *Control Eng. Pract.* **2021**, 110.
(27) Jiang, Q.; Yan, X.; Huang, B. Neighborhood Variational Bayesian Multivariate Analysis for Distributed Process Monitoring With Missing Data. *IEEE Trans. Control Syst. Technol.* **2019**, *27* (6), 2330−2339.
(28) Sefidian, A. M.; Daneshpour, N. Missing Value Imputation Using a Novel Grey Based Fuzzy C-Means, Mutual Information Based Feature Selection, and Regression Model. *Expert Syst. Appl.* **2019**, *115*, 68−94.
(29) Chen, X.; Wei, Z.; Li, Z.; Liang, J.; Cai, Y.; Zhang, B. Ensemble Correlation-Based Low-Rank Matrix Completion with Applications to Traffic Data Imputation. *Knowl. based Syst.* **2017**, *132*, 249−262.
(30) Ku, W.; Storer, R. H.; Georgakis, C. Disturbance Detection and Isolation by Dynamic Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1995**, *30* (1), 179−196.
(31) Yao, L.; Ge, Z. Industrial Big Data Modeling and Monitoring Framework for Plant-Wide Processes. *IEEE Trans. Ind. informatics* **2021**, *17*, 6399.
(32) Yoon, J.; Jordon, J.; van der Schaar, M. GAIN: Missing Data Imputation Using Generative Adversarial Nets. *International Conference on Machine Learning* **2018**, 5675−5684.
(33) Lin, W.-C.; Tsai, C.-F. Missing Value Imputation: A Review and Analysis of the Literature (2006−2017). *Artif. Intell. Rev.* **2020**, *53* (2), 1487−1509.
(34) Yu, W.; Zhao, C. Low-Rank Characteristic and Temporal Correlation Analytics for Incipient Industrial Fault Detection with Missing Data. *IEEE Trans. Ind. informatics* **2021**, *17*, 6337−6346.
(35) Yao, L.; Ge, Z. Moving Window Adaptive Soft Sensor for State Shifting Process Based on Weighted Supervised Latent Factor Analysis. *Control Eng. Pract.* **2017**, *61*, 72−80.
(36) Zhang, X.; Li, Y.; Kano, M. Quality Prediction in Complex Batch Processes with Just-in-Time Learning Model Based on Non-Gaussian Dissimilarity Measure. *Ind. Eng. Chem. Res.* **2015**, *54* (31), 7694−7705.
(37) Zhao, C.; Wang, W.; Tian, C.; Sun, Y. Fine-Scale Modeling and Monitoring of Wide-Range Nonstationary Batch Processes With Dynamic Analytics. *ieee Trans. Ind. Electron.* **2021**, *68* (9), 8808−8818.
(38) Kaneko, H.; Funatsu, K. A New Process Variable and Dynamics Selection Method Based on a Genetic Algorithm-based Wavelength Selection Method. *AIChE J.* **2012**, *58* (6), 1829−1840.
(39) Chatfield, C. The Analysis of Time Series: An Introduction, Sixth Edition, 2017.
(40) Tong, C.; Shi, X. Decentralized Monitoring of Dynamic Processes Based on Dynamic Feature Selection and Informative Fault Pattern Dissimilarity. *IEEE Trans. Ind. Electron.* **2016**, *63* (6), 3804−3814.
(41) Wang, A.; Chen, Y.; An, N.; Yang, J.; Li, L.; Jiang, L. Microarray Missing Value Imputation: A Regularized Local Learning Method. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2019**, *16* (3), 980−993.
(42) Jiang, Q.; Yan, S.; Cheng, H.; Yan, X. Local−Global Modeling and Distributed Computing Framework for Nonlinear Plant-Wide Process Monitoring With Industrial Big Data. *IEEE Trans. neural networks* **2021**, *32* (8), 3355−3365.
(43) Chiang, L. H.; Russell, E. L.; Braatz, R. D. Fault Detection and Diagnosis in Industrial Systems. *Meas. Sci. Technol.* **2001**, *12* (10), 1745.
(44) Little, R.; Rubin, D. Statistical Analysis with Missing Data, Second Edition. 2019. DOI: 10.1002/9781119482260.
(45) Walczak, B.; Massart, D. L. Dealing with Missing Data: Part II. *Chemom. Intell. Lab. Syst.* **2001**, *58* (1), 29−42.
(46) Severson, K. A.; Molaro, M. C.; Braatz, R. D. Principal Component Analysis of Process Datasets with Missing Values. *Processes* **2017**, *5*, 38.
(47) Yuan, X.; Ge, Z.; Huang, B.; Song, Z. A Probabilistic Just-in-Time Learning Framework for Soft Sensor Development With Missing Data. *IEEE Trans. Control Syst. Technol.* **2017**, *25* (3), 1124−1132.
(48) Liu, Y.; Yang, C.; Zhang, M.; Dai, Y.; Yao, Y. Development of Adversarial Transfer Learning Soft Sensor for Multigrade Processes. *Ind. Eng. Chem. Res.* **2020**, *59*, 16330.
(49) Negiz, A.; Çlinar, A. Statistical Monitoring of Multivariable Dynamic Processes with State-Space Models. *AIChE J.* **1997**, *43* (8), 2002−2020.
(50) Zhong, L.; Chang, Y.; Wang, F.; Gao, S. Distributed Operating Performance Assessment of the Plant-Wide Process Based on Data-Driven Hybrid Characteristics Decomposition. *Ind. Eng. Chem. Res.* **2020**, *59* (35), 15682−15696.
(51) Chang, Y.; Ma, R.; Wang, F.; Zheng, W.; Wang, S. Multimode Process Mode Identification With Coexistence of Quantitative Information and Qualitative Information. *IEEE Trans. Autom. Sci. Eng.* **2020**, *17*, 1516−1527.

https://doi.org/10.1021/acs.iecr.1c03860
Ind. Eng. Chem. Res. 2022, 61, 580−593