I&EC research
Industrial & Engineering Chemistry Research

Article

pubs.acs.org/IECR

Vine Copula-Based Dependence Description for Multivariate Multimode Process Monitoring

Xiang Ren,[†] Ying Tian,[‡,†] and Shaojun Li*[,†]

[†]Key Laboratory of Advanced Control and Optimization for Chemical Processes, Ministry of Education, East China University of Science and Technology, Shanghai 200237, China

[‡]School of Optical-electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

Supporting Information

ABSTRACT: A novel vine copula-based dependence description (VCDD) process monitoring approach is proposed. The main contribution is to extract the complex dependence among process variables rather than perform dimensionality reduction or other decoupling processes. For a multimode chemical process, the C-vine copula model of each mode is initially created, in which a multivariate optimization problem is simplified as coping with a series of bivariate copulas listed in a sparse matrix. To measure the distance of the process data from each non-Gaussian mode, a generalized local probability (GLP) index is defined. Consequently, the generalized Bayesian inference-based probability (GBIP) index under a given control limit can be further calculated in real time via searching the density quantile table created offline. The validity and effectiveness of the proposed approach are illustrated using a numerical example and the Tennessee Eastman benchmark process. The results show that the proposed VCDD approach achieves good performance in both monitoring results and computation load.

1. INTRODUCTION

In modern industries, risk management to handle abnormal events has been an active area of research. An abnormal event occurs due to a series of factors such as process unit aging, control system failure, external disturbances and noise, etc., which directly results in process variables departing from the normal ranges. If timely action is not taken, an abnormal event will deteriorate from an incipient fault to a near-miss, to an incident, and finally an accident will occur. To maintain process safety and product quality, the first step is to detect and monitor variations from a deluge of process variables and thus avoid further propagation of the abnormal events.[1]

Methods for process monitoring can be divided into three categories: model-based methods,[1] knowledge-based methods,[2] and data-based methods.[3] In recent years, due to the existence of large scale database and the development of the associated techniques that are efficient enough to handle the big data, data-based process monitoring methods become more and more popular.[4,5] Data-based methods have no requirement of the exact model and the associated expert knowledge, among which principal component analysis (PCA)[6] and partial least-squares (PLS)[7] have been widely investigated. These two multivariate statistical process monitoring (MSPM) methods, however, have some limitation in assumptions that the process data are Gaussian-distributed and the relationships among process variables are linear-correlated. However, when the data are non-Gaussian and highly nonlinear, the assumptions are no longer satisfied. In this respect, other data-based methods have been introduced: Kernel-based method,[8] neural network,[9] manifold learning,[10] etc. perform well in tackling nonlinear problems, while independent component analysis,[11] the Gaussian mixture model,[12] support vector data description,[13] etc. are good at describing non-Gaussian processes. Meanwhile, various improvements and combinations of the existing MSPM methods[14−18] also have been proposed, achieving a better monitoring performance for the complex industrial processes under certain conditions.

In the modeling process of fault detection, usually, the training data are labeled as "normal" or "anomaly". When identifying the anomalies is of great cost or the knowledge of the process data is limited or unreliable, the widespread supervised approaches for process monitoring may appear invalid. Recently, Escobar et al.[19] proposed an unsupervised approach for nonlinear process monitoring. Multidimensional data were mapped to a low-dimensional latent space grid using a generative topographical mapping technique. Important data information described by probability distribution was then analyzed through similarity assessment, and finally a kind of graph clustering method was performed to fulfill fault identification. This method, with a focus on exploiting the correlation relationships among variables, achieves comparable performance with other supervised methods. Generally, the current monitoring approaches (supervised or unsupervised) tend to simplify the problem (high dimensionality) by projecting data onto a lower dimensional subspace and then further analysis would be made. It seems that dimensionality reduction or the decoupling process is a natural or even an exclusive way in MSPM. Therefore, it is of great value to provide an efficient and relatively less time-consuming process monitoring strategy by describing the intricate dependencies among variables directly.

As an efficient statistical tool in dependence modeling, copula has become increasingly popular in many fields of application, particularly in economics,[20,21] finance,[22,23] and

Received: April 3, 2015
Revised: September 1, 2015
Accepted: September 27, 2015

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

meteorology.[24,25] In recent years, copula has received much attention in chemical process systems engineering. Seider and co-workers[26−28] applied Cuadras-Augé copula and multivariate Gaussian copula to describe a nonlinear relationship between variables and behavior-based factors involving human operators in chemical process risk analysis. Ahooyi et al.[29] proposed a moment-based method for estimating the probabilities of a rare event, and a copula-based model was illustrated as a comparison study. However, the copula theory is scarcely applied in the field of process monitoring at present, which is mainly due to its cumbersome and inefficient optimization procedure of traditional multivariate copulas,[30] known as "curse of dimensionality".

To overcome such a problem, vines are proposed and developed by analyzing various dependence structures.[31−34] Vines are graphical models with bivariate copulas as building blocks of the joint distribution. The most attracting property of vines is that they simplify multivariate dependence problems into solving a series of optimization problems of bivariate copulas, which can avoid a high level of computational cost for the rather high dimensional variables. In addition, vine copula is able to consider both nonlinearity and non-Gaussian behaviors of the process data; note that the monitoring performance will improve when considering different data characteristics as much as possible (enough data information can be preserved when modeling). The advantages of vine copula are 2-fold: (1) It is able to capture high nonlinearity behavior by using a rank correlation coefficient such as Kendall and Spearman rank correlation coefficient as a dependence measurement (Unlike linear correlation coefficient such as Pearson correlation coefficient, rank correlation coefficient is able to capture more than the first or the second moment information on the sample data.); (2) Due to the flexible structure of vines, the corresponding copula model has many more abilities in describing a multitude of non-Gaussian processes (various combinations of different marginal distributions and copulas), especially those with heavy tail dependence and asymmetry properties.

In this article, a novel vine copula-based dependence description (VCDD) monitoring approach is proposed. The main issue of the proposed approach is, from the probabilistic view, how to analyze, capture, and model the complex dependence and interaction among variables. Another concern of the VCDD approach is how to define and calculate a kind of monitoring index applicable for arbitrary distributions. Note that the widely used Mahalanobis distance is no longer efficient in measuring the distance of data from each non-Gaussian mode. Referring to the work, C-vine copula, one type of vine, is initially employed to establish the joint distribution of each mode, which can efficiently catch the complex dependencies among high dimensional variables. The Markov chain Monte Carlo (MCMC) method with the Metropolis-Hastings (M-H) algorithm (a sampling method) is performed to obtain various statistical information on each mode. The Bayesian inference-based probability (BIP) index proposed by Yu and Qin[35] is improved to achieve multimode process monitoring and a generalized local probability (GLP) index is defined. Herein, the highest density region (HDR) of arbitrary distributions (a mathematical terminology that has been defined in ref 36) is employed, which is found to be a good measurement of the GLP index. To calculate the GLP index with less computation load, the density quantile approach (DQA) is introduced. More specifically, the joint probability density function (PDF) values in the normal operating region are calculated and discretized according to different confidence levels, thus, a brief density quantile table with finite numbers of intervals is then created, ensuring that the GLP index of the real-time data, as well as the corresponding generalized Bayesian inference-based probability (GBIP) index, can immediately be updated by just searching the static table created offline.

The rest of the article is organized as follows. A detailed dependence modeling strategy with C-vine copula is given in Section 2, which includes pair copula construction, conditional distribution calculation, selection and optimization of vine copula models. In Section 3, a kind of C-VCDD monitoring approach is further developed, along with a brief description of DQA. By regarding the PCA-based and finite Gaussian mixture model (FGMM) approach[35] as comparison, the validity and effectiveness of the proposed approach are illustrated through two application examples: a numerical example (Section 4) and the Tennessee Eastman (TE) benchmark process (Section 5). Finally, Section 6 presents some conclusions.

2. DEPENDENCE MODELING WITH VINE COPULA

In this section, we first give a brief review of copula and vine copula. A detailed description on the structure and simulation of C-vine is then provided in the following three subsections.

A copula is a multivariate distribution function whose margins are all uniformed over (0,1). It can be used to provide multivariate dependence structure separately from the margins and build a bridge between a multivariate cumulative distribution and the corresponding univariate cumulative marginal distributions. According to the Sklar theorem,[37] an n-dimensional joint distribution can be decomposed into its n univariate marginal distributions and an n-dimensional copula. Mathematically, let F be the n-dimensional joint cumulative distribution function (CDF) of the random vector $\mathbf{x} = [x_1,x_2,...,x_n]^T$, then there exists a copula function C such that

$$F(\mathbf{x}) = C(F_1(x_1), F_2(x_2), ..., F_n(x_n)) \tag{1}$$

where $F_i(x_i)$ $(i = 1,2,...,n)$ denotes the ith marginal CDF satisfying

$$F_i(x_i) = u_i = \int_{-\infty}^{x_i} f_i(\bar{x}_i)\mathrm{d}\bar{x}_i, \qquad u_i \in [0, 1] \tag{2}$$

Given that C is differentiable, take the derivative of $\mathbf{x}$ in eq 1, the corresponding joint PDF, $f(\mathbf{x})$ can be obtained as a product of the copula density c and the marginal PDF $f_1f_2,...f_n$

$$f(\mathbf{x}) = c(u_1, u_2, ..., u_n)\prod_{i=1}^{n} f_i(x_i) \tag{3}$$

where c is defined as

$$c(u_1, u_2, ..., u_n) = \frac{\partial^n C}{\partial u_1 \partial u_2 ... \partial u_n} \tag{4}$$

Since several copula structures are available to describe different dependence behaviors, the process of copula modeling, actually, is to use the sample data to fit the best copula in eq 4. Once the structure and the parameter of the copula are determined and estimated, the optimized joint PDF of the multidimensional data then can be obtained in eq 3. (More details about copula optimization are discussed in Subsection 2.3.) Obviously, the optimization problem that fits the sample data to copula density in eq 4 gets harder when the number of parameters to be estimated in eq 4 increases, thus resulting in

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

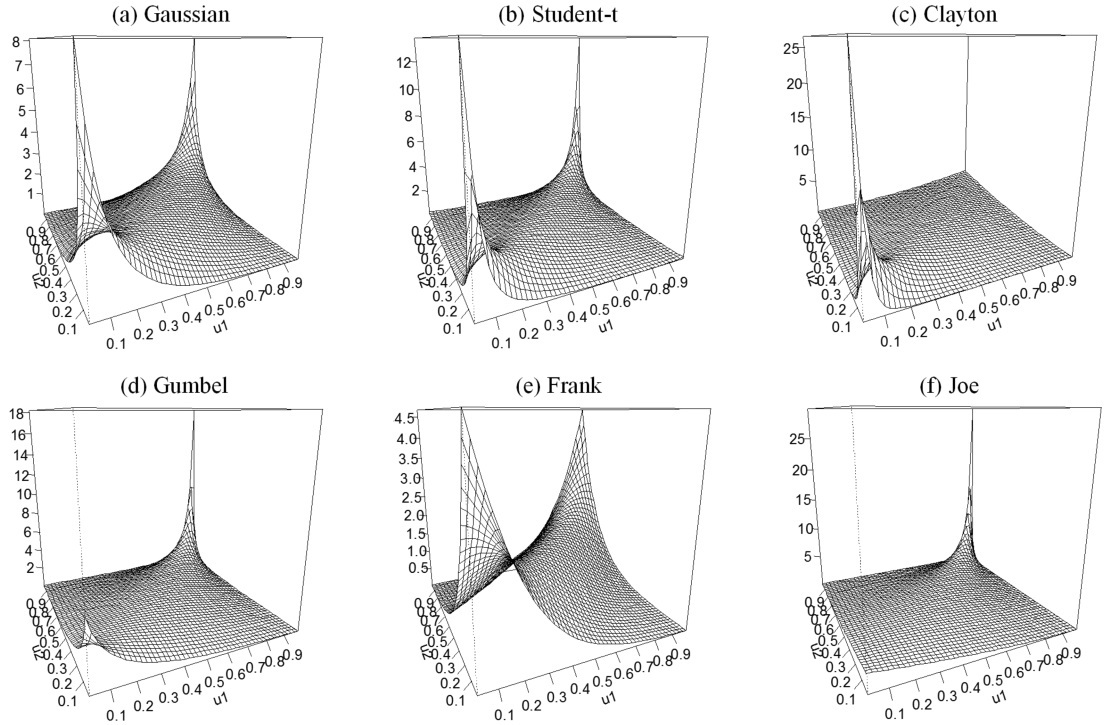

Figure 1. PDF diagrams of six typical bivariate copulas with Kendall rank correlation coefficients all set as 0.5.

an expensive computation load for conventional multivariate copulas.

Unlike some conventional multivariate copulas, e.g., multivariate Gaussian copula, that apply the same dependence structure to each pair of variables, vine copula exhibits much more flexibility by invoking abundant dependence information from various types of bivariate copulas. Figure 1 displays six typical bivariate copulas, involving two Elliptical copulas: Gaussian and Student-t and four Archimedean copulas: Clayton, Gumbel, Frank, and Joe. These bivariate copulas, controlled by different numbers of parameters (usually one or two), are able to capture different types of tail dependence behaviors. For example, Clayton with one constricted parameter allows for nonzero lower tail dependence coefficient, while Student-t with two constricted parameters has both nonzero lower and upper tail dependence coefficients. Denotation and properties of each bivariate copula can be seen in ref 30. Different combinations of structure and parameter of bivariate copulas enable vine copula to describe a wide range of processes with complex dependence behaviors. However, it should be noted that the number of the available bivariate copulas is finite, and it is impossible to describe all kinds of the correlation information completely. Therefore, it is in essential to select a best-fit copula rather than the true data-generating one, and goodness-of-fit testing is needed. Generally, bivariate copulas to be selected could be predetermined and simplified using some prior information. For example, if one pair of variables reflects strong upper tail dependence and little lower tail dependence, then those with zero upper tail dependence can be excluded from the predefined set. This procedure may easily help search an appropriate bivariate copula as soon as possible.

2.1. Pair Copula Construction. Vine copula provides a theoretical way to decompose multivariate copula into a series of bivariate copulas, termed as pair copulas. A number of factorization forms of vine copulas are available, among which C-vine and D-vine are the most typical ones, called regular vines.[38] In this work, we mainly focus on the factorization strategy of C-vine, theoretically and graphically.

With respect to a random vector $\mathbf{x}$ with length n, if $x_1,x_2,...,x_n$ are exchangeable, then the decomposition of its joint PDF can be formulated as

$$f(\mathbf{x}) = f(x_1)\prod_{t=2}^{n} f(x_t|x_1, x_2, ..., x_{t-1}) \tag{5}$$

To further analyze each conditional PDF in eq 5, herein, consider the joint PDF of $x_{t-i}(i = 1,2,...,t-1)$ and x_t conditioned on $x_1,x_2,...,x_{t-i-1}$, i.e., $f(x_{t-i},x_t|x_1,x_2,...,x_{t-i-1})$. According to the property of conditional distribution and the Sklar theorem,[37] we have that

$$\begin{aligned} &f(x_{t-i}, x_t|x_1, x_2, ..., x_{t-i-1}) \\ &\quad = f(x_{t-i}|x_1, x_2, ..., x_{t-i-1})\cdot f(x_t|x_1, x_2, ..., x_{t-i-1}, x_{t-i}) \\ &\quad = c_{t-i,t|1:t-i-1}(F(x_{t-i}|x_1, x_2, ..., x_{t-i-1}), F(x_t|x_1, x_2, ..., x_{t-i-1})) \\ &\quad \times f(x_{t-i}|x_1, x_2, ..., x_{t-i-1})\cdot f(x_t|x_1, x_2, ..., x_{t-i-1}) \end{aligned} \tag{6}$$

Therefore,

$$\begin{aligned} &f(x_t|x_1, x_2, ..., x_{t-i-1}, x_{t-i}) \\ &\quad = c_{t-i,t|1:t-i-1}(F(x_{t-i}|x_1, x_2, ..., x_{t-i-1}), F(x_t|x_1, x_2, ..., x_{t-i-1})) \\ &\quad \times f(x_t|x_1, x_2, ..., x_{t-i-1}) \end{aligned} \tag{7}$$

where $c_{t-i,t|1:t-i-1}$ denotes the bivariate copula density, $f(x_t|x_1,x_2,...,x_{t-i+1},x_{t-i})$ is the conditional PDF of x_t given $x_1,x_2,...,x_{t-i}$, $F(x_t|x_1,x_2,...,x_{t-i-1})$ represents the conditional CDF of x_t. Eq 7 demonstrates that a conditional PDF with long-length conditional variables can be simplified as that with shorter-length conditional variables, hence, repeated application with

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

i = 1,2,...,t−1 results in the following decomposition of each conditional PDF in eq 5

$$f(x_t|x_1, x_2, ..., x_{t-1}) = f(x_t)\prod_{i=1}^{t-1} c_{t-i,t|1:t-i-1}$$
$$(F(x_{t-i}|x_1, x_2, ..., x_{t-i-1}), F(x_t|x_1, x_2, ..., x_{t-i-1})) \quad (8)$$

where $c_{t-i,t|1:t-i-1}(F(x_{t-i}|x_1,x_2,...,x_{t-i-1}),F(x_t|x_1,x_2,...,x_{t-i-1})) = c_{1,t}(F(x_1),F(x_t))$ when $i = t-1$. Eq 8 depicts the decomposition strategy of C-vine, one of the most popular vines. It is able to describe multivariate copulas using a rich variety of bivariate copulas as building blocks, so-called pair copulas, whose modeling scheme is based on the decomposition of an n-dimensional multivariate density into $n(n-1)/2$ bivariate copula densities.

For ease of understanding, take the C-vine model in five dimensions as an example. A star structure of C-vine trees (acyclic connected graphs with nodes and edges) is given in Figure 2. It is shown that a five-dimensional C-vine copula

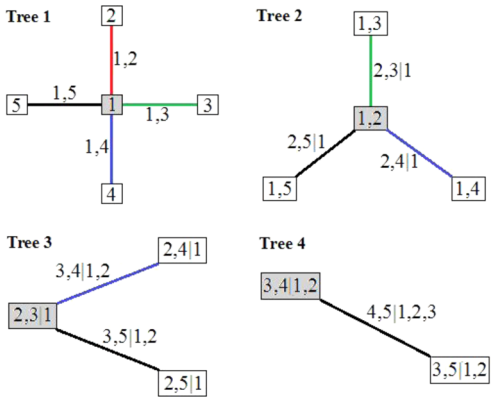

Figure 2. Illustration of five-dimensional C-vine trees with edge indices (corresponding to pair copulas).

consists of four linked trees along with a total of ten pair copulas (bivariate copulas). Each tree has a root node, and the root node of the current tree is chosen from the first edge of the previous tree. In the first tree, the predefined root node is x_1 (determining the root node is an optimization problem that will be further discussed in Subsection 2.3), and it directly causes relationships with x_2,x_3,x_4,x_5, respectively. Four edges represent the corresponding four correlated relationships, which can be mathematically expressed as four pair copulas, $c_{1,2}$,$c_{1,3}$,$c_{1,4}$,$c_{1,5}$. Based on the first tree, then it comes to the second one. The root node in the second tree has no exact meaning, it is just an abstract sign that helps us identify the pair copulas in the second tree, i.e., $c_{2,3|1}$,$c_{2,4|1}$,$c_{2,5|1}$. Similarly, other trees follow the same procedure that all pairwise dependencies with respect to the root node in a specific tree are modeled conditioned on all previous nodes. It is worth noting that all the pair copulas except those in the first tree include conditional distribution functions. For example, $c_{1,3}$ in the first tree is short for $c_{1,3}(F(x_1),F(x_3);\theta_{1,3})$, while $c_{2,5|1}$ in the second tree is short for $c_{2,5|1}(F(x_2|x_1),F(x_5|x_1);\theta_{2,5|1})$. $\theta_{1,3}$, the parameter of pair copula, $c_{1,3}$ can be easily optimized on the basis of the empirical CDF values of x_1 and x_3. However, in order to estimate $\theta_{2,5|1}$, we have to have a deep understanding of the conditional distributions, $F(x_2|x_1)$ and $F(x_5|x_1)$. (Calculation of conditional distribution functions in each tree and estimation of parameter in each pair copula will be further described in Subsection 2.2 and Subsection 2.3.) Overall, the C-vine copula model in five dimensions or the joint CDF of five-dimensional random variables is expressed as

$$f(x_1, x_2, x_3, x_4, x_5) = \prod_{t=1}^{5} f(x_t)$$
$$\times c_{1,2}c_{1,3}c_{1,4}c_{1,5}c_{2,3|1}c_{2,4|1}c_{2,5|1}c_{3,4|1,2}c_{3,5|1,2}c_{4,5|1,2,3} \quad (9)$$

Based on the graphical analysis above, a more generalized C-vine copula model is essentially a combination of pair copulas in the whole trees, and its standard density formula with arbitrary dimensions is given by

$$f(\mathbf{x}) = \prod_{t=1}^{n} f_t(x_t) \times \prod_{i=1}^{n-1}\prod_{j=1}^{n-i} c_{i,i+j|1:i-1}(F(x_i|x_1, ..., x_{i-1}),$$
$$F(x_{i+j}|x_1, ..., x_{i-1}); \theta_{i,i+j|1:i-1}) \quad (10)$$

where $\theta_{i,i+j|1:i-1}$ denotes the copula parameter to be estimated between conditional CDFs of x_i and x_{i+j} both given x_1,x_2,...,x_{i-1}. It should be emphasized that eq 10 is a graph-induced expression of C-vine, which actually equals the aforementioned analytical expression, i.e., eq 5 combined with eq 8. The difference is that the graph-induced expression specifies the pair copulas tree by tree (easy to understand), while the analytical expression is strictly proved and runs over the pairs with the order of different color edges in Figure 2.

2.2. Calculating Conditional Distribution Function. As has been mentioned in Subsection 2.1, a challenging problem for the inference of vine copula is how to evaluate the conditional distribution functions in eq 10. To solve this problem, various analytical expressions of bivariate copulas have been studied, along with the general notion of the h-function proposed by Aas et al.,[34] which is given by

$$h_{x_i,x_j|\tilde{\mathbf{x}}}\left(F(x_i|\tilde{\mathbf{x}})|F(x_j|\tilde{\mathbf{x}}); \theta_{x_i,x_j|\tilde{\mathbf{x}}}\right) \triangleq F(x_i|x_j, \tilde{\mathbf{x}}) = \frac{\partial C_{x_i,x_j|\tilde{\mathbf{x}}}(F(x_i|\tilde{\mathbf{x}}), F(x_j|\tilde{\mathbf{x}}); \theta_{x_i,x_j|\tilde{\mathbf{x}}})}{\partial F(x_j|\tilde{\mathbf{x}})} \quad (11)$$

where x_i, x_j are scalars, $\tilde{\mathbf{x}}$ is a column vector excluding x_i and x_j, which also satisfies $\{\mathbf{x}\} = \{x_i,x_j,\tilde{\mathbf{x}}\}$, and $\{\mathbf{x}\}$ represents the data set consisting of all elements in $\mathbf{x}$; $F(x_i|\tilde{\mathbf{x}})$ denotes the conditional CDF of x_i given $\tilde{\mathbf{x}}$; and $C_{xi,xj|\tilde{\mathbf{x}}}$ denotes the copula function corresponding to its pair copula density $c_{xi,xj|\tilde{\mathbf{x}}}$ between $F(x_i|\tilde{\mathbf{x}})$ and $F(x_j|\tilde{\mathbf{x}})$ with the related parameter $\theta_{xi,xj|\tilde{\mathbf{x}}}$. For better understanding, a brief demonstration of eq 11 is provided as follows.

Note that

$$f(x_i|x_j, \tilde{\mathbf{x}}) = \frac{f(x_i, x_j|\tilde{\mathbf{x}})}{f(x_j|\tilde{\mathbf{x}})} \quad (12)$$

We have

$$F(x_i|x_j, \tilde{\mathbf{x}}) = \int_{-\infty}^{x_i} f(z_i|x_j, \tilde{\mathbf{x}})\mathrm{d}z_i = \int_{-\infty}^{x_i} \frac{\partial^2}{\partial z_i \partial x_j} F(z_i, x_j|\tilde{\mathbf{x}})\mathrm{d}z_i \cdot \frac{1}{f(x_j|\tilde{\mathbf{x}})} \quad (13)$$

According to the Sklar theorem[37]

$$F(z_i, x_j|\tilde{\mathbf{x}}) = C_{z_i,x_j|\tilde{\mathbf{x}}}(F(z_i|\tilde{\mathbf{x}}), F(x_j|\tilde{\mathbf{x}}); \theta_{z_i,x_j|\tilde{\mathbf{x}}}) \quad (14)$$

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

So that

$$F(x_i|x_j, \tilde{\mathbf{x}}) = \int_{-\infty}^{x_i} \frac{\partial^2}{\partial z_i \partial x_j} C_{z_i,x_j|\tilde{\mathbf{x}}}(F(z_i|\tilde{\mathbf{x}}), F(x_j|\tilde{\mathbf{x}}); \theta_{z_i,x_j|\tilde{\mathbf{x}}}) \mathrm{d}z_i \cdot \frac{1}{f(x_j|\tilde{\mathbf{x}})}$$
$$= \frac{\partial C_{x_i,x_j|\tilde{\mathbf{x}}}(F(x_i|\tilde{\mathbf{x}}), F(x_j|\tilde{\mathbf{x}}); \theta_{x_i,x_j|\tilde{\mathbf{x}}})}{\partial F(x_j|\tilde{\mathbf{x}})} \cdot \frac{\partial F(x_j|\tilde{\mathbf{x}})}{\partial x_j} \cdot \frac{1}{f(x_j|\tilde{\mathbf{x}})}$$
$$= \frac{\partial C_{x_i,x_j|\tilde{\mathbf{x}}}(F(x_i|\tilde{\mathbf{x}}), F(x_j|\tilde{\mathbf{x}}); \theta_{x_i,x_j|\tilde{\mathbf{x}}})}{\partial F(x_j|\tilde{\mathbf{x}})} \tag{15}$$

Based on the *h*-function analyzed above, all of the conditional distribution functions in a C-vine model can be computed through an iterative way. As is shown more clearly in Figure 3,

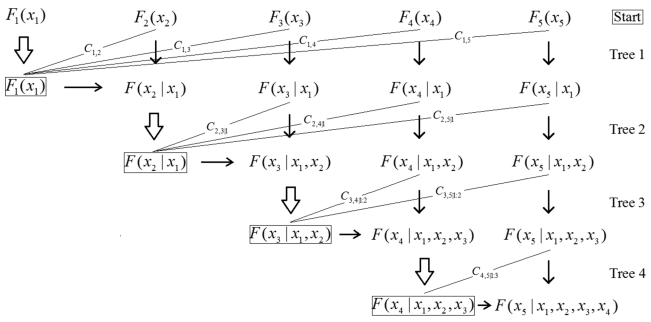

Figure 3. An iterative strategy for calculating conditional distribution functions, where C-vine trees are also embedded.

five layers with conditional distribution functions are established. On the basis of the given (empirical) marginal CDFs, $F_i(x_i)$ (i = 1,2,...,5) in the first layer, the available conditional distribution functions are calculated through eq 11 layer by layer. The first term of the conditional distribution function in the current layer is a common element for calculating all of the conditional distribution functions in the next layer. For example, the conditional distribution function $F(x_3|x_1,x_2)$ in the third layer is calculated using $F(x_2|x_1)$ and $F(x_3|x_1)$ in the second layer

$$F(x_3|x_1, x_2) = \frac{\partial C_{2,3|1}(F(x_2|x_1), F(x_3|x_1); \theta_{2,3|1})}{\partial F(x_2|x_1)} \tag{16}$$

while $F(x_4|x_1,x_2)$ is computed using $F(x_2|x_1)$ and $F(x_4|x_1)$

$$F(x_4|x_1, x_2) = \frac{\partial C_{2,4|1}(F(x_2|x_1), F(x_4|x_1); \theta_{2,4|1})}{\partial F(x_2|x_1)} \tag{17}$$

It is necessary to emphasize that this layer-by-layer calculation process cannot be fulfilled at one time. For example, if we want to calculate $F(x_3|x_1,x_2)$ in eq 16, both the conditional distribution function values in the previous layer, $F(x_2|x_1)$, $F(x_3|x_1)$, and the optimized parameter of the copula function $C_{2,3|1}$ in the second tree, $\theta_{2,3|1}$ (see Figure 2), should be known in advance. Therefore, the process of conditional distribution calculation and parameter estimation should be performed in turn, i.e., start with the empirical CDF values of each variable in the first layer → optimize parameter of each pair copula in the first tree → calculate all conditional distribution function values in the second layer → optimize parameter of each pair copula in the second tree →...→ calculate all conditional distribution function values in the last layer. (Parameter estimation for each pair copula will be further discussed in Subsection 2.3.)

2.3. Selection and Optimization of Vine Copula Models. Due to different variable order combinations, actually, a rich variety of C-vine models are available. For example, in Figure 2, there are five candidates (x_1,x_2,x_3,x_4,x_5) to be chosen as the root node in the first tree, indicating that at least five different C-vine models are available. To select a relatively appropriate C-vine model, Aas et al.[34] pointed out that it was preferable to consider the pair copulas with high dependence, which could lead to smaller quantities of pair copulas to be optimized. In this work, we choose the $\hat{i}$th variable as the root node in the first C-vine tree by maximizing the following equation

$$\hat{i} = \arg\max_i \sum_{j=1}^{n} |\tau_{i,j}| \quad (i = 1, 2, ..., n) \tag{18}$$

where $\tau_{i,j}$ represents the Kendall rank correlation coefficient of random variables X_i and X_j, which is defined as[39]

$$\tau_{i,j} = \Pr((X_i - \tilde{X}_i)(X_j - \tilde{X}_j) > 0) - \Pr((X_i - \tilde{X}_i)(X_j - \tilde{X}_j) < 0) \tag{19}$$

where $(\tilde{X}_i,\tilde{X}_j)$ is an independent copy of (X_i,X_j). Each probability term in eq 19 can be estimated from its frequency count of observations. In ref 39, it points out that, compared with the linear correlation coefficient such as the Pearson correlation coefficient, the Kendall rank correlation coefficient is invariant under strictly increasing transformations and is the correct dependence measure to use for other joint distributions rather than multivariate Gaussian distribution only, which is found to be an efficient coefficient that associates with copula. Let $\tau_{i,j} = 1$ when $i = j$. The rest of the nodes in the first tree are then sorted with the decreasing values of $\tau_{\hat{i},j}$. In particular, if $\sum_{j=1}^{n}|\tau_{\hat{i},j}| = \sum_{j=1}^{n}|\tau_{i_1,j}| = \sum_{j=1}^{n}|\tau_{i_2,j}|$ and $i_1 < i_2$, then variable i_1 is set as the root node in the first tree; if $\tau_{\hat{i},j_1} = \tau_{\hat{i},j_2}$ and $j_1 < j_2$, then we give priority to variable j_1, ensuring that the node orders in the overall trees are fixed. For example, assuming that the Kendall rank correlation coefficient among three random variables are $\tau_{1,2}$ = 0.4, $\tau_{1,3}$ = 0.2, and $\tau_{2,3}$ = 0.9, according to eq 18, the objective function values are 1.6, 2.3, and 2.1 with different root nodes selected. x_2 is considered as the best root node because it achieves the largest value (2.3). Note that $\tau_{1,2} < \tau_{2,3}$, and the variable order selected is x_2,x_3,x_1.

Indeed, there exist large numbers of possible decompositions of vines especially for high dimensional variables. Though a proper structure can be selected from a diversity of C-vine copula models using eq 18, it is actually not always the best one among all decompositions. Usually, it is suggested to consider all possible decompositions for smaller dimensions ($n \leq 4$), i.e., selection of a specific factorization. However, this procedure becomes rather cumbersome when the dimension increases rapidly. In this condition, the decomposition could be determined roughly according to statistical properties of sample data, e.g., focusing on the bivariate relationships that are most important to model correctly.[34] In this paper, we aim to present a detailed description of C-vine that is able to achieve relatively good results, as is shown in Section 4 and Section 5. A more general strategy coping with all vine factorizations in process monitoring requires further study.

Once the C-vine model is established, we can proceed with the estimation of copula parameters via a fully parametric maximum likelihood estimation (MLE). Now suppose

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

M observations $\mathbf{X}_k$ (k = 1,2,...,M) are available to fit the fully parametric density, combined with eq 3, we have that

$$f_{\alpha,\beta}(\mathbf{x}) = c(F_{1,\beta(1)}(x_1), F_{2,\beta(2)}(x_2), ..., F_{n,\beta(n)}(x_n); \alpha) \prod_{t=1}^{n} f_{t,\beta(t)}(x_t) \tag{20}$$

where α denotes the overall parameters and family orders of C-vine copula, and β denotes the parameters of the margins, $\beta = \{\beta(1),\beta(2),...,\beta(n)\}$. By taking logarithms, the fully parametric MLE paradigm tries to estimate

$$\begin{aligned}(\hat{\alpha}, \hat{\beta}) &= \arg\max_{\alpha,\beta} \int \log f_{\alpha,\beta}(\mathbf{x}) \mathrm{d}F(\mathbf{x}) \\ &= \arg\max_{\alpha,\beta} \left\{ \lim_{M\to\infty} \frac{1}{M} \sum_{k=1}^{M} \log f_{\alpha,\beta}(\mathbf{X}_k) \right\}\end{aligned} \tag{21}$$

Eq 21 provides a generalized fitting method to arbitrary copulas. It has also been proved that the fully parametric MLE method asymptotically minimizes the Kullback−Leibler divergence (relative entropy) between a true data-generating copula and an available parametric copula.[38] However, to simplify the optimization problem, it is suggested to obtain the parameters of margins previously and select each bivariate copula (structure and parameter) separately. The parametric distribution (e.g., Gaussian, Chi Square distributions, etc.) estimation and the kernel method provide an efficient tool in describing data behavior to model univariate marginal distributions,[40] and a vine copula model selection formula, termed as maximum pseudolikelihood-based Akaike information criterion (MPL-AIC), is described as follows

$$\begin{aligned}&(\hat{\theta}_{i,i+j|1:i-1}, \hat{\gamma}_{i,i+j|1:i-1}) \\ &= \arg\max_{\substack{\theta_{i,i+j|1:i-1} \\ \gamma_{i,i+j|1:i-1}}} \left\{ \sum_{k=1}^{M} \log[c(F^k(x_i|x_1, ..., x_{i-1}), \right. \\ &\left. F^k(x_{i+j}|x_1, ..., x_{i-1}); \theta_{i,i+j|1:i-1}, \gamma_{i,i+j|1:i-1})] - \lambda \right\}\end{aligned} \tag{22}$$

where $\theta_{i,i+j|1:i-1}$, $\gamma_{i,i+j|1:i-1}$ denote the parameter and the family order of the corresponding pair copula, respectively. $F^k(x_i|x_1,...,x_{i-1})$, $F^k(x_{i+j}|x_1,...,x_{i-1})$ denote the kth observation of $F(x_i|x_1,...,x_{i-1})$, $F(x_{i+j}|x_1,...,x_{i-1})$ respectively, which can be calculated with a recursive form in eq 11. λ denotes the number of parameters to be estimated for a specific bivariate copula, in most cases, λ = 1,2.

The vine copula optimization procedure is basically described as follows: (1) Choose a specific bivariate copula from a predefined set; in other words, the family order of the current pair copula, $\gamma_{i,i+j|1:i-1}$ is fixed. (2) Estimate and record the parameter of the current pair copula $\theta_{i,i+j|1:i-1}$ that maximizes eq 22. (3) Repeat step (1) and step (2) for all of the bivariate copulas in the predefined set. (4) Select the best copula from the predefined set that achieves the largest value of eq 22. (5) Repeat step (1)−step (4) tree by tree, as shown in Figure 2. It is worth noting that the parameter of each bivariate copula has its particular range, hence, the aforementioned procedure is essentially to solve constraint optimization problems. For example, if the current pair copula from the predefined set is Clayton, whose CDF is given for $\theta > 0$ by

$$C(u_1, u_2; \theta) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-1/\theta} \tag{23}$$

and the corresponding copula density function is

$$c(u_1, u_2; \theta) = (1 + \theta)(u_1u_2)^{-\theta-1}(u_1^{-\theta} + u_2^{-\theta} - 1)^{-2-1/\theta} \tag{24}$$

where θ denotes the parameter of Clayton, and $u_i = F(x_i)$ represents the marginal distribution of random variable X_i. Given that γ = 3 (fixed), λ = 1 (Clayton density has only one parameter to be estimated), and the associated parameter constraint $\theta > 0$, according to eq 22, the constraint optimization problem is

$$\begin{aligned}&\max \quad \sum_{k=1}^{M} \log[(1 + \theta)(u_1^k u_2^k)^{-\theta-1}(u_1^{k-\theta} + u_2^{k-\theta} - 1)^{-2-1/\theta}] - 1 \\ &\text{s.t.} \quad \theta > 0\end{aligned} \tag{25}$$

where u_1^k denotes the kth observation of u_1. The L-BFGS-B algorithm can be applied to solve the constraint optimization problem easily. For one-parameter bivariate copula families (λ = 1), a method called inversion of Kendall's τ (Kendall rank correlation coefficient) can be used to set the starting values for optimization, e.g., the relationship between Kendall's τ and the estimated copula parameter θ of Clayton is $\tau = \theta/(\theta+2)$. For more details, see ref 41.

The CDVine package[42] for the statistical software R is available for the simulation of C-vine and D-vine. It provides 32 bivariate copulas, including the rotated versions (rotated by 90, 180, 270 deg) of Clayton (γ = 3), Gumbel (γ = 4), Joe (γ = 6), and four BB families (γ = 7,8,9,10), as the fitted copulas which can describe a wide range of dependence behaviors. Note that what we discuss here is the best-fitted copula available rather than the correct copula (true data-generating copula), hence, goodness-of-fit testing is usually needed.

Based on the analysis mentioned above and combined with the iterative schematic given in Figure 2 and Figure 3, the overall optimization procedures for a C-vine copula model can be summarized as follows:

(1) Select the optimum bivariate copulas (e.g., among the 32 bivariate copulas in the predefined set) to fit the original data in tree 1; both parameter estimation and structure optimization are included.

a. For a specific bivariate copula, i.e., the copula family order $\gamma_{i,i+j|1:i-1}$ is fixed, estimate its corresponding parameter(s) using eq 22. This estimation is easy to perform, because only two (or only one) dimension(s) are (is) handled at a time;
b. Choose the pair $(\theta_{i,i+j|1:i-1}, \gamma_{i,i+j|1:i-1})$ that achieves the largest value in eq 22 as the optimum bivariate copula.

(2) With respect to tree i(i = 2,...,n−1), compute the implied observations (i.e., conditional distribution function values) using the optimized pair copulas from tree i−1 and the appropriate h-function discussed in eq 11.

(3) Select the optimum bivariate copulas in tree i. This procedure is similar to step (1) except that the original data are replaced by the implied observations that have been calculated in step (2).

(4) Perform $i = i + 1$. Stop when $i = n$; otherwise, go back to step (2).

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

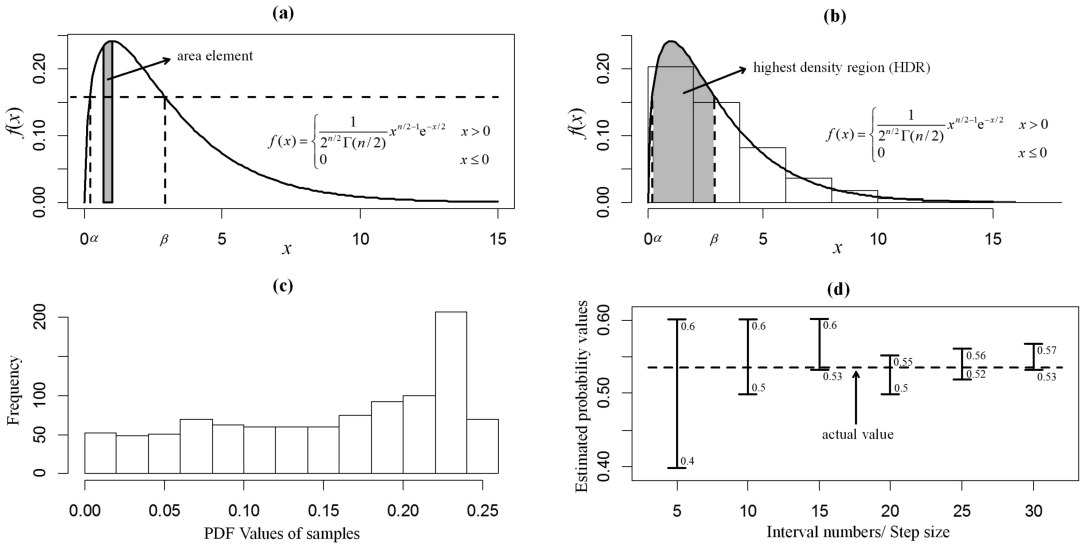

Figure 4. Schematic of the numerical integration approach and the density quantile approach to calculate probability in HDR.

3. C-VCDD MULTIVARIATE MULTIMODE PROCESS MONITORING

Multimode processes are often encountered in industrial production. In this section, a kind of vine copula-based approach is proposed to fulfill multivariate multimode process monitoring. Before that, an efficient mathematical technique called the density quantile approach (DQA) is discussed in Subsection 3.1, which is helpful in estimating the generalized local probability (GLP) index used for updating the generalized Bayesian inference-based probability (GBIP) index.

3.1. Probability Calculation Using the Density Quantile Approach. For a given distribution function, the main issue is to calculate the probability value in a specific region. Such a region, termed as the highest density region (HDR), is found to be a valid probability-based distance measurement applicable for arbitrary distribution, on the premise that the associated PDF is known. Hyndman has provided a precise definition of HDR as follows.[36]

Definition. Let $f(x)$ be the PDF of a random variable X. Then the $1-\delta$ HDR is the subset $R(f_\delta)$ of the sample space of X such that

$$R(f_\delta) = \{x|f(x) \geq f_\delta\} \tag{26}$$

where f_δ is the largest constant that $\Pr(X \in R(f_\delta)) \geq 1-\delta$, $1-\delta$ represents the confidence level (also a probability value) of HDR.

Our main task, however, is to find the confidence level, in other words, estimate the probability value of the specific HDR in a given distribution. Generally, there are two approaches fulfilling such estimation, i.e., the numerical integration approach and the density quantile approach. Without lose of generality, further explanation is made using an example with univariate distribution.

Figure 4(a) depicts the PDF diagram of Chi Square distribution (non-Gaussian) with degree of freedom set as 3. The HDR discussed here is $R(f_\delta) = [\alpha,\beta]$, where $f(\alpha) = f(\beta)$. To calculate the probability value in HDR, one direct approach is to divide the HDR into a series of area elements and integrate over the HDR, that is

$$\Pr(X \in R(f_\delta)) = \Pr(\alpha \leq X \leq \beta) = \int_\alpha^\beta f(x)\mathrm{d}x \tag{27}$$

Eq 27 can be estimated using the numerical integral approach. However, this method has two serious defects: (1) a great deal of computation complexity for high dimensional variables (solving a multiple integral problem) and (2) requiring complete information on integral domain (both the exact values of α and β should be known).

To handle the aforementioned problems, the density quantile approach is proposed. From a statistical view, DQA is focused on obtaining the information on the probability coverage of a given region of the sample space via a Monte Carlo technique. According to the definition of HDR, it is found that the probability value, $1-\delta$, also represents the confidence level of the HDR. Eq 26 shows that the confidence level of the HDR is related to the quantile of the random variable $f(X)$. Therefore, the issue of probability estimation is equal to searching its corresponding quantile. Consider the random variable $Y = f(X)$, according to the definition of HDR, it is shown that f_δ satisfies

$$\Pr(\alpha \leq X \leq \beta) = \Pr(f(X) \geq f_\delta) = 1 - \delta \tag{28}$$

where f_δ is the δ quantile of Y, which can be estimated as the $\hat{\delta}$ sample quantile from the sample set of Y. The sample set of Y is denoted by $\mathbf{Y} = \{Y_i|i = 1,2,...,M\}$, where M denotes the number of samples/observations in $\mathbf{Y}$, $Y_i = f(X_i)$, X_i denotes the ith observations sampled from the PDF, $f(x)$. In essence, we just use the $\hat{\delta}$ sample quantile of $\mathbf{Y}$, denoted by $q_{\hat{\delta}}^{\mathbf{Y}}$, as an estimate of the δ quantile of Y, denoted by f_δ. For example, $\mathbf{Y}$ is sampled from the distribution of the random variable Y. Assume $\mathbf{Y} = \{2,2,1,5,4,4,4\}$, we initially sort the whole observations in the ascending order, i.e., $\tilde{\mathbf{Y}} = \{1,2,2,4,4,4,5\}$, then calculate the cumulative frequency values of different observations $\{1/7,3/7,3/7,6/7,6/7,6/7,1\}$, so $q_{\hat{\delta}}^{\mathbf{Y}} = 1,2,4,5$ is the $\hat{\delta} = (1/7,3/7,6/7,1)$ sample quantile of $\mathbf{Y}$, respectively. If the number of the observations in the sample set is large enough, it is suitable to conclude that the estimated $\delta = (1/7,3/6,6/7,1)$ quantile of Y is $f_\delta = 1,2,4,5$, respectively. Once the sample quantiles under different confidence levels are obtained, the connection between the probability value of the HDR and the quantiles then can be further discussed. By comparing the PDF values of

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 1. Density Quantile Approach: Statistical Information on PDF Values with Interval Step Size Set as 5

parameter	value					
confidence level	0	0.2	0.4	0.6	0.8	1
density quantile	0.0005	0.0735	0.1398	0.1944	0.2291	0.2420

the boundary, $f(\alpha)$ and the quantiles of $Y = f(X)$ previously obtained, we can easily estimate the probability value in the HDR, $R(f_\delta) = [\alpha,\beta]$ through a particular confidence level interval.

The observations from arbitrary distribution, $f(x)$, are obtained using a sampling method, e.g., the MCMC method. To calculate the sample quantiles from the observations, the number of the confidence levels should be predefined, which is controlled by the interval step size l. Figure 4(b)-(d) clearly shows the procedure and the performance of DQA. In Figure 4(b), 1000 observations X_i (i = 1,2,...,1000) are sampled from the Chi Square distribution by which the density histogram is provided. Calculate $Y_i = f(X_i)$ by transforming X_i by its own density function, then another 1000 observations Y_i (i = 1,2,...,1000), termed as PDF values, are obtained. The corresponding histogram is depicted in Figure 4(c). Table 1 shows the statistical information (confidence level and sample quantile) of Y_i (i = 1,2,...,1000) on the condition that l = 5. The boundary of the HDR discussed here is assumed as α = 0.2. Given that $f(0.2)$ = 0.1614, satisfying $f(\alpha) \in (q^{\mathbf{Y}}_{0.4}, q^{\mathbf{Y}}_{0.6})$ = (0.1398, 0.1944), we have that

$$\Pr(X \in R(f_\delta)) = \Pr(0.2 \le X \le \beta) \in (1 - 0.6, 1 - 0.4) = (0.4, 0.6) \tag{29}$$

Indeed, such an estimation can be made arbitrarily accurate by increasing the number of observations, M, and the interval step size, l. Figure 4(d) depicts the probability interval estimation with six interval step sizes. It shows that the range of the estimation interval for estimating the actual value turns narrowed for a larger interval step size.

Note that DQA is able to identify HDR by just analyzing the quantiles of the one-dimensional PDF values rather than the distribution of multiple random variables, and this method achieves a much lower computation load especially for rather high dimensional variables, which is found to be an efficient tool for MSPM.

3.2. Construction and Estimation of the GBIP Index.

For the n-dimensional process data, the joint PDF of the kth (k = 1,2,...,K) mode C_k, denoted by $f^{(k)}(\mathbf{x})$, is first established using C-vine copula, as shown in eq 10. It is worth noting that the prior information on the training data belonging to their real operating modes should be known in advance, otherwise a data clustering method is suggested. Then Bayesian inference strategy proposed by Yu and Qin[35] is introduced, and a generalized monitoring index for non-Gaussian distribution, whose control limit is denoted by $CL \in [0,1]$, is defined as

$$\mathrm{GBIP} = \sum_{k=1}^{K} P(C_k|\mathbf{X}_t^{\mathrm{monitor}}) P_L^{(k)}(\mathbf{X}_t^{\mathrm{monitor}}) \tag{30}$$

$P(C_k|\mathbf{X}_t^{\mathrm{monitor}})$ represents the posterior probability of the monitored data $\mathbf{X}_t^{\mathrm{monitor}}$ belonging to $f^{(k)}(\mathbf{X})$ and can be computed through eq 31

$$P(C_k|\mathbf{X}_t^{\mathrm{monitor}}) = \frac{P(C_k) f^{(k)}(\mathbf{X}_t^{\mathrm{monitor}})}{\sum_{i=1}^{K} P(C_i) f^{(i)}(\mathbf{X}_t^{\mathrm{monitor}})} \quad (k = 1, 2, \ldots, K) \tag{31}$$

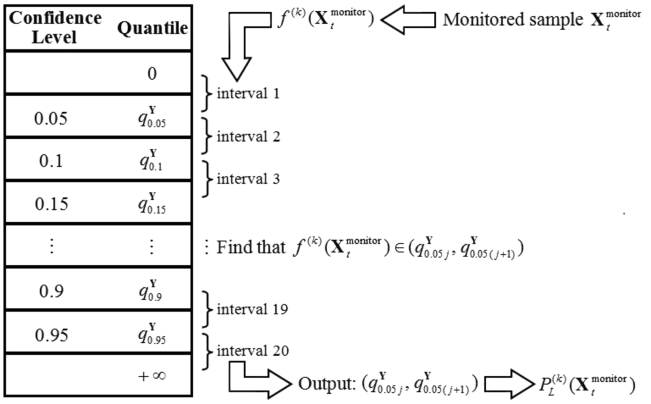

Figure 5. Density quantile table: A searching schematic for estimating the GLP index interval of monitored samples.

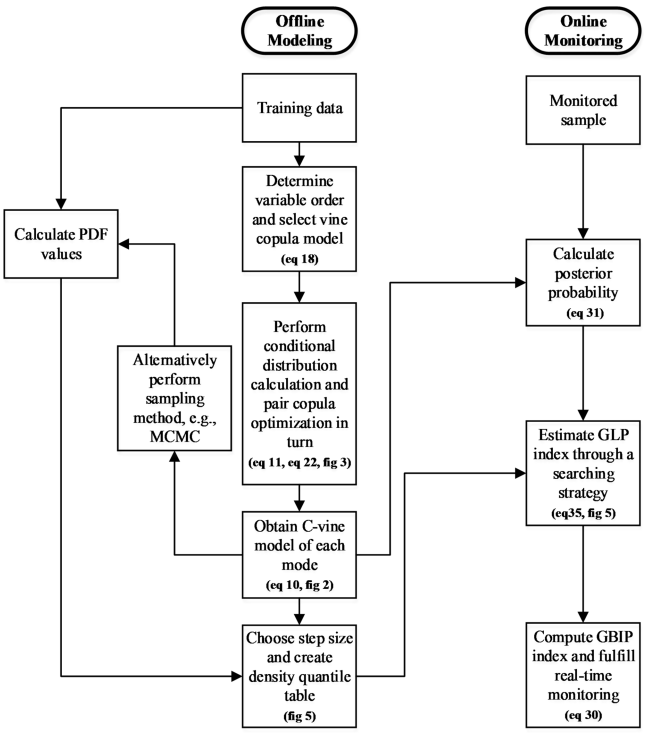

Figure 6. Algorithm flowchart of the C-VCDD monitoring approach.

where $P(C_k)$ is the prior probability of the monitored data belonging to $f^{(k)}(\mathbf{x})$. The noninformative prior or the prior information extracted from the previous monitored data in a time window is suggested. More practically, if n_k (k = 1,2,...,K) training data are clustered into each of the K modes, respectively, then we can simply set

$$P(C_k) = \frac{n_k}{\sum_{i=1}^{K} n_i} \quad (k = 1, 2, \ldots, K) \tag{32}$$

$P_L^{(k)}(\mathbf{X}_t^{\mathrm{monitor}})$ represents the GLP index of the data to each non-Gaussian mode. Herein, HDR is employed to describe

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

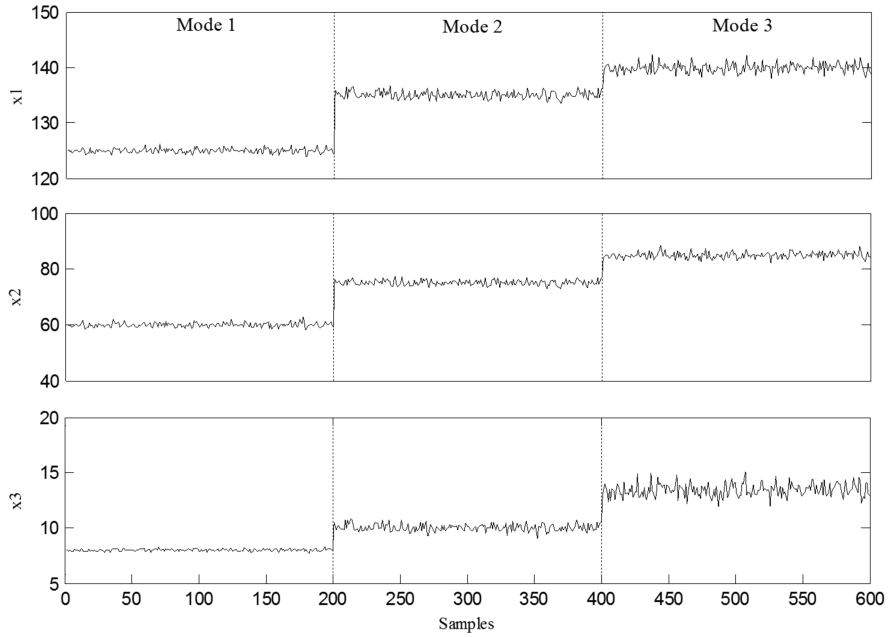

Figure 7. Numerical example: Totally 600 training data in three dimensions under three operating modes.

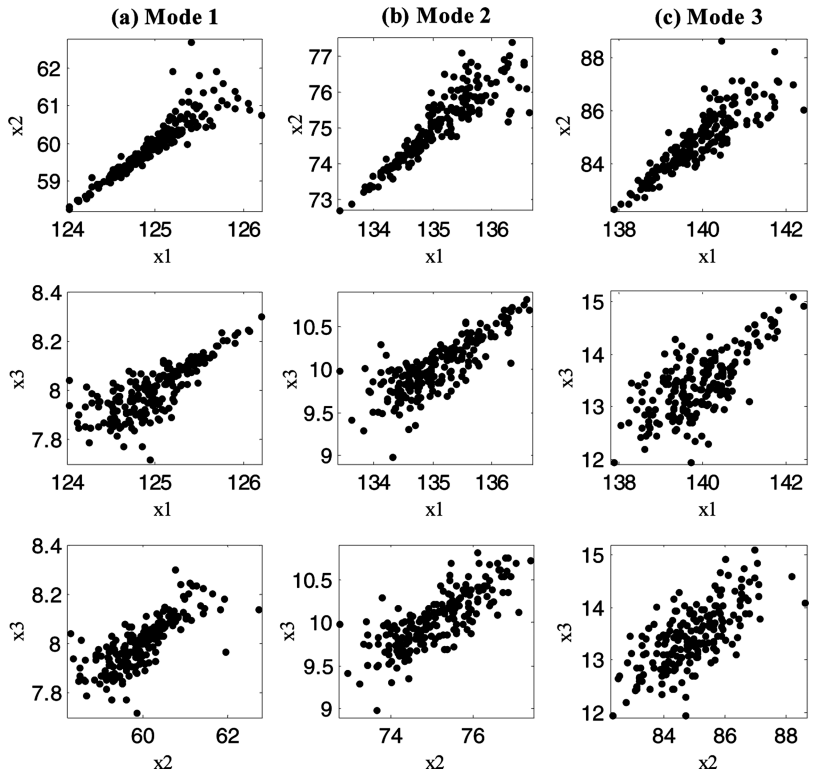

Figure 8. Numerical example: Scatter plots of the training data, each column corresponds to one mode.

$P_L^{(k)}(\mathbf{X}_t^{\text{monitor}})$ instead of the Mahalanobis distance that performs well in each Gaussian component only. According to Subsection 3.1, the GLP index relative to the kth mode can be defined as

$$P_L^{(k)}(\mathbf{X}_t^{\text{monitor}}) = \Pr(f^{(k)}(\mathbf{X}) \geq f^{(k)}(\mathbf{X}_t^{\text{monitor}})) \tag{33}$$

where $\mathbf{X}$ is a random vector. $f^{(k)}(\mathbf{X})$, the joint PDF of mode k, can also be treated as a mapping function whose domain of definition is $\mathbf{X}$. Let $\mathbf{X}_i$ $(i = 1,2,...,M)$ be the samples from distribution $f^{(k)}(\mathbf{X})$, Y_i denote the corresponding joint PDF value, where $Y_i = f^{(k)}(\mathbf{X}_i)$ and $\mathbf{Y} = \{Y_i | i = 1,2,...,M\}$. For the current monitored sample $\mathbf{X}_i^{\text{monitor}}$, if it satisfies

$$q_{\hat{\delta}}^{\mathbf{Y}} = f^{(k)}(\mathbf{X}_t^{\text{monitor}}) \tag{34}$$

then

$$1 - \hat{\delta} \xrightarrow[M\to\infty]{a.s} P_L^{(k)}(\mathbf{X}_t^{\text{monitor}}) \tag{35}$$

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 2. Numerical Example: Pair Copula Optimization for Three Modes

	mode 1			mode 2			mode 3		
pair copula	$c_{1,2}$	$c_{1,3}$	$c_{2,3\|1}$	$c_{1,2}$	$c_{1,3}$	$c_{2,3\|1}$	$c_{1,2}$	$c_{1,3}$	$c_{2,3\|1}$
Kendall's tau	0.8	0.6	0.4	0.75	0.55	0.3	0.7	0.5	0.2
copula order	3	6	5	3	6	1	3	4	1
copula type	Clayton	Joe	Frank	Clayton	Joe	Gaussian	Clayton	Gumbel	Gaussian
parameter 1	8	3.827	4.161	6	3.286	0.454	4.667	2	0.309
parameter 2	\	\	\	\	\	\	\	\	\

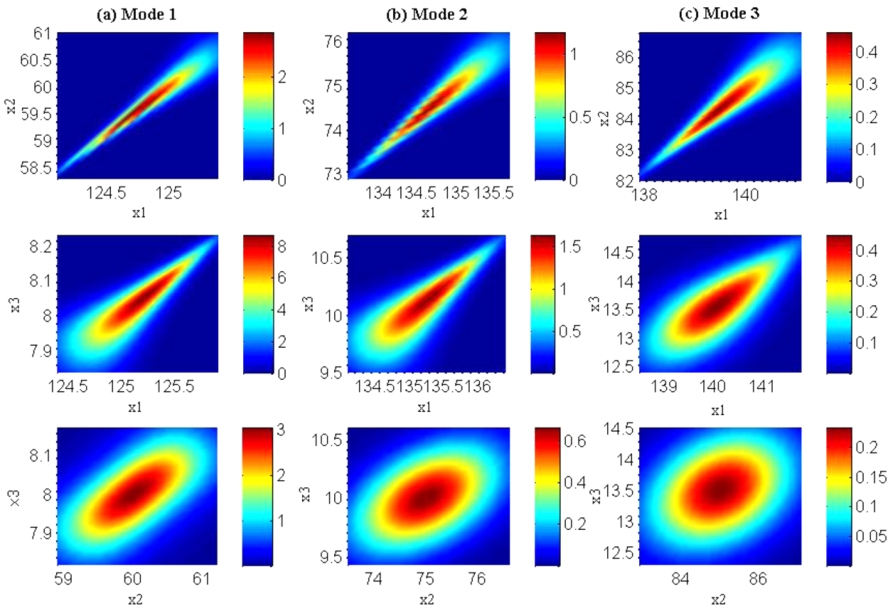

Figure 9. Numerical example: Distributions of two-dimensional variables for the training data, each column corresponds to one mode.

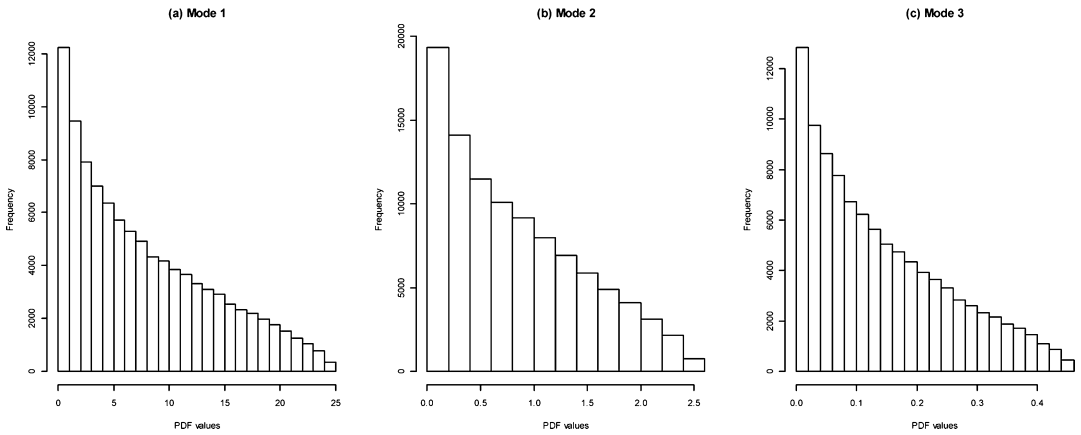

Figure 10. Numerical example: Histogram of joint PDF values for each mode.

where $\hat{\delta}$ represents the estimated confidence level corresponding to a given quantile value of $\mathbf{Y}$, $q_{\hat{\delta}}^{\mathbf{Y}}$. According to eqs 33-35, it is shown that the created GLP index corresponds to the related confidence level of the density quantile. Actually, the GLP index is an extended form of the local probability index defined in ref 35 that uses the Mahalanobis distance as a measurement for Gaussian distribution only.

For computation simplification, a density quantile table is designed to estimate the GLP index in eq 35. The confidence levels of $\mathbf{Y}$ in a discrete form are considered. Define l as the step size for discretization, then a brief table with l intervals (including l+1 confidence levels and the corresponding l+1 density quantile values), termed as the density quantile table, is created. It should be noted that, on the one hand, a smaller l leads to a more accurate estimation but may somewhat increase time in estimating GLP via a searching strategy; on the other hand, the step size should be selected according to the control limit, so that each discrete interval (confidence levels) obtained does not go across the control limit, hence, it will not interfere with our judgment whether the exact GBIP index exceeds the control limit or not. After making a compromise, the step size is suggested to set as $l = (1/1-CL)$ for a given control limit CL. A more general form of step size chosen criterion is $l = (AC/1-CL)$, where AC denotes

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 3. Numerical Example: Density Quantile Table to Fulfill a Searching Strategy, Three Modes Are Included

	density quantile		
confidence level	mode 1	mode 2	mode 3
0	0	0	0
0.05	0.343	0.041	0.007
0.1	0.780	0.091	0.015
0.15	1.279	0.146	0.024
0.2	1.810	0.208	0.035
0.25	2.405	0.275	0.045
0.3	3.057	0.348	0.057
0.35	3.767	0.425	0.07
0.4	4.522	0.511	0.083
0.45	5.343	0.602	0.098
0.5	6.245	0.699	0.114
0.55	7.212	0.800	0.131
0.6	8.249	0.905	0.15
0.65	9.428	1.019	0.17
0.7	10.669	1.144	0.192
0.75	12.020	1.280	0.217
0.8	13.547	1.429	0.244
0.85	15.283	1.603	0.277
0.9	17.410	1.806	0.317
0.95	19.991	2.064	0.367
1	24.829	2.549	0.46

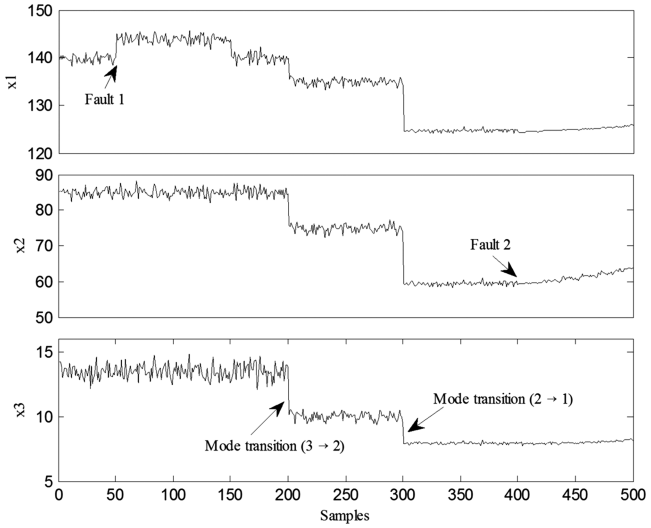

Figure 11. Numerical example: Time series plot of the 500 monitored data in each dimension.

accuracy coefficient and satisfies AC = 1,2,3,.... For example, on the condition that CL = 0.95, if the step size is set as l = 10 (AC = 0.5), then a density quantile table with 10 intervals is obtained. Now assume that the exact value of the GLP index is 0.92, by searching the corresponding density quantile table, it is probably found that $P_L^{(k)}(\mathbf{X}_t^{\text{monitor}}) \in (0.9, 1)$. Unfortunately, we are still unable to decide whether the GLP index exceeds 0.95 due to its too large estimated interval. If l = 20 (AC = 1), the searching result $P_L^{(k)}(\mathbf{X}_t^{\text{monitor}}) \in (0.9, 0.95)$ obviously decreases the estimated range of GLP. If l = 40 (AC = 2), we can find that $P_L^{(k)}(\mathbf{X}_t^{\text{monitor}}) \in (0.9, 0.925)$, and this will lead to a more accurate estimation of GLP and GBIP index values. When the step size is large enough, it is appropriate to set the value of the GLP index as the average of its upper and lower bounds, i.e., $P_L^{(k)}(\mathbf{X}_t^{\text{monitor}}) = 0.925$ with AC = 1 or $P_L^{(k)}(\mathbf{X}_t^{\text{monitor}}) = 0.9125$ with AC = 2. In this way, the GLP index is estimated as a point rather than an interval.

In this work, we select l = 20 with the control limit set as CL = 95%, and a density quantile table with 20 intervals is created, which includes $[0, q^{Y}_{0.05})$, $[q^{Y}_{0.05j}, q^{Y}_{0.05(j+1)})$ (j = 1,2,...,18), and $[q^{Y}_{0.95}, +\infty)$. Figure 5 depicts a searching schematic for estimating the interval of the GLP index. For the current monitored sample $\mathbf{X}_t^{\text{monitor}}$, its joint PDF value in each mode, $f^{(k)}(\mathbf{X}_t^{\text{monitor}})$, should be calculated first. By searching the static table created offline, the interval satisfying $f(\mathbf{X}_t^{\text{monitor}}) \in (q^{Y}_{0.05j}, q^{Y}_{0.05(j+1)})$ is identified, and the corresponding interval of the GLP index, accordingly, can be estimated through eq 35.

The samples used for creating the density quantile table can be sampled by the MCMC method with the M-H algorithm. An associated description of the MCMC method for the D-vine model has been given in ref 43. However, in many application cases, there is no need to generate the samples using the MCMC method when the training data are large enough since those data are capable of reflecting the whole information on each mode. Therefore, the available training data can be used to obtain the joint PDF values, which can significantly reduce the computation load. Overall, the algorithm flowchart of the proposed VCDD approach is shown in Figure 6, and the detailed procedures are given as follows.

Offline Modeling. (1) Collect a set of historical training data under all possible modes. The information on data belonging to the real operating modes can be obtained via expert knowledge or performing clustering methods.

(2) Establish a C-vine model for each mode, which includes (a) determining variable orders; (b) constructing C-vine models; (c) calculating conditional distribution functions and optimizing pair copulas layer by layer, tree by tree; and (d) performing goodness-of-fit testing.

(3) Obtain the samples of the C-vine model for each mode using the MCMC method. This step can be eliminated when the historical training data are large enough to cover most information on each operating mode.

(4) Specify a control limit CL, choose the step size for discretization according to $l = (1/1-CL)$, then estimate the discrete density quantiles with different confidence levels, and create the corresponding static density quantile table

Online Monitoring. (1) Calculate the posterior probability of the monitored sample $\mathbf{X}_t^{\text{monitor}}$ belonging to each operating mode, $P(C_k|\mathbf{X}_t^{\text{monitor}})$ using eq 31. The noninformative prior or the prior information extracted from the previous monitored data in a time window is suggested.

(2) Compute the joint PDF values of $\mathbf{X}_t^{\text{monitor}}$ under all modes (C-vine models) and estimate the GLP index of each mode through a searching strategy.

(3) Estimate the GBIP index through eq 30. According to the given control limit CL in the GBIP control chart, detect the abnormal operating conditions for $\mathbf{X}_t^{\text{monitor}}$ satisfying GBIP > CL.

4. A NUMERICAL EXAMPLE

In this section, a numerical example is discussed to validate the efficiency of the proposed C-VCDD approach in handling a non-Gaussian process with heavy tail dependence. The simulated data are supplied in the Supporting Information. Three operating modes are analyzed, each of which has 200 training samples in three dimensions. Figure 7 displays the time series plots of the sample data, and the corresponding scatter

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

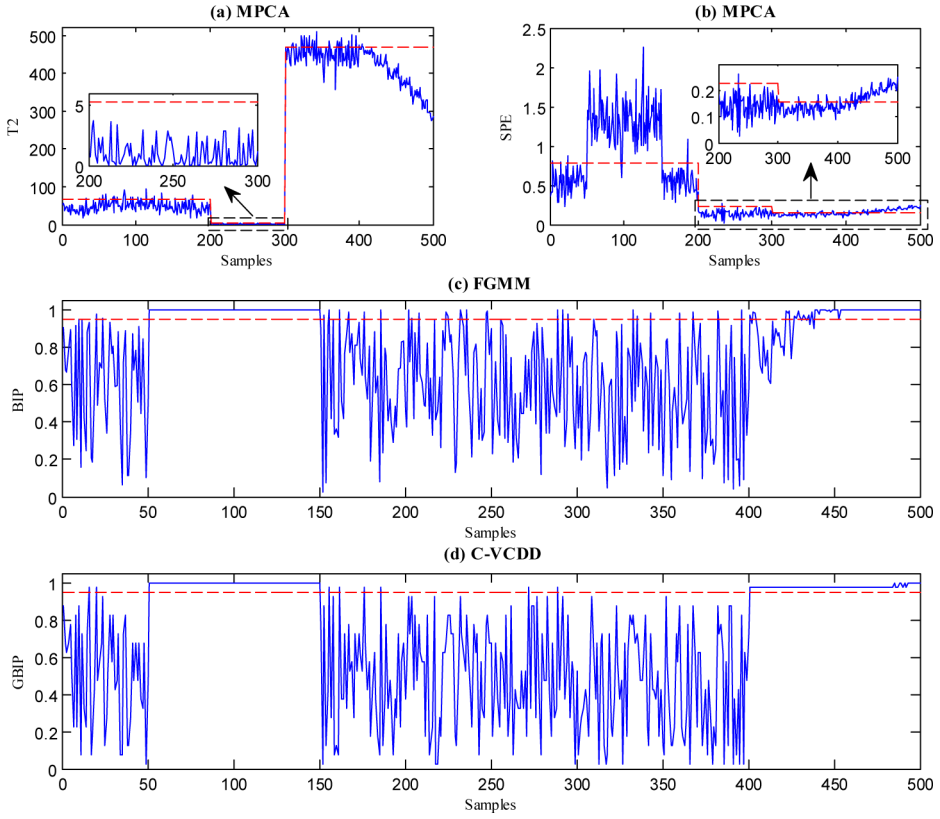

Figure 12. Numerical example: Monitoring control charts with 95% control limit based on (a) MPCA-T^2, (b) MPCA-SPE, (c) FGMM-BIP, and (d) C-VCDD-GBIP.

Table 4. Numerical Example: Monitoring Performance Analysis in a Comparison Study

		MPCA		FGMM	C-VCDD
samples studied	evaluation index	T^2	SPE	BIP	GBIP
Fault 1 (100 samples)	MDR	0.81	0.01	0	0
Fault 2 (100 samples)	MDR	0.96	0.26	0.25	0
	DT	100	38	38	0
Non-Fault (300 samples)	FDR	0.1067	0.0433	0.0933	0.0267

plots for the three modes are displayed in Figure 8. It shows that there exists heavy lower tail dependence between x_1 and x_2 and upper tail dependence between x_1 and x_3, while tail dependence between x_2 and x_3 is not that significant. Herein, the marginal distribution of the training data in each mode can be simply assumed as Gaussian distribution through hypothesis testing, which makes the process of C-vine copula modeling more convenient.

In offline modeling, 200 training data for each mode are used to establish the C-vine model, the selected variable order is obtained using eq 18 (variables are reordered as x_1,x_2,x_3), and all pair copula families and parameters in each tree are estimated through eq 22, as shown in Table 2. For visualization, distributions of two-dimensional variables created by bivariate copula models are depicted in Figure 9, and it shows that the optimized C-vine models describe the non-Gaussian distributions well, which is mainly due to the flexible structure of vines. For example, mode 1 involves three different types of bivariate copulas, Clayton, Joe, and Frank. Clayton can describe lower tail dependence behavior; Joe, higher tail dependence behavior (see Figure 1). Obviously, less types of bivariate copulas to be chosen will lead to limited ability or even poor performance in capturing the complex dependence behaviors. Note that the multivariate Gaussian copula has no ability to describe tail dependence behavior, and it is difficult for the Gaussian mixture model to establish such an accurate model by invoking a finite number of Gaussian components.

To obtain the density quantile table of the process analyzed, 100000 joint PDF values of the three C-vine models are calculated using MCMC simulation, whose histograms are given in Figure 10. It is noticed that the distribution of joint PDF values exhibits a crucial difference compared to that of the multivariate Gaussian distribution in the FGMM approach.[35] Combined with the DQA discussed in Subsection 3.1, the density quantile table of the numerical example is obtained in Table 3, where density quantile under confidence level $\delta = 0$, q_0^Y, is also estimated.

The monitored data within 500 time periods are set as follows: the process initially runs at mode 3 for the first 50 time periods, then a bias error of 4 is added to x_1, after 100 time periods going by, it returns to the normal operating condition; in time periods 201−300, the system works at mode 2; then switches to mode 1 for another 100 time periods, afterward, a bias error of 0.5 along with a nonlinear drifting error is added to x_2. The time series plot of the 500 monitored samples is shown in Figure 11. It is worth noting that Fault 2 which is mainly caused by the drifting error is much less visible than Fault 1, so that in many occasions we are likely to regard it as a normal state especially in time periods 401−440. Yet, it is indeed a typical type of fault that reflects significant dependence

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

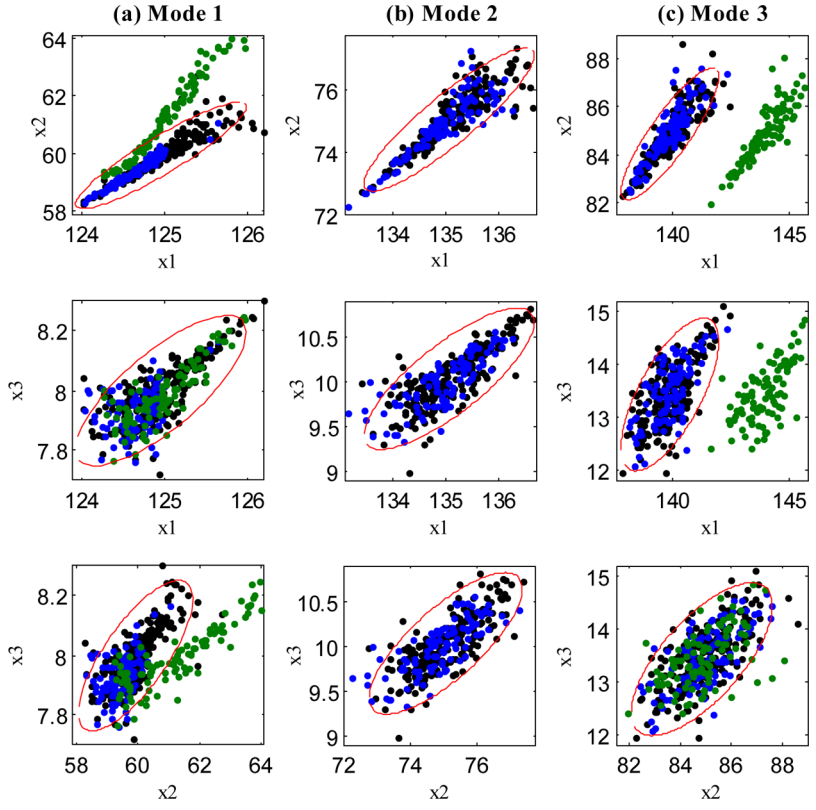

Figure 13. Numerical example: Scatter plots of training data (black points) and monitored data — the faulty samples are represented by green points, while those under normal operating conditions are represented by blue ones. The estimated 95% probability ellipses based on the FGMM approach are also provided.

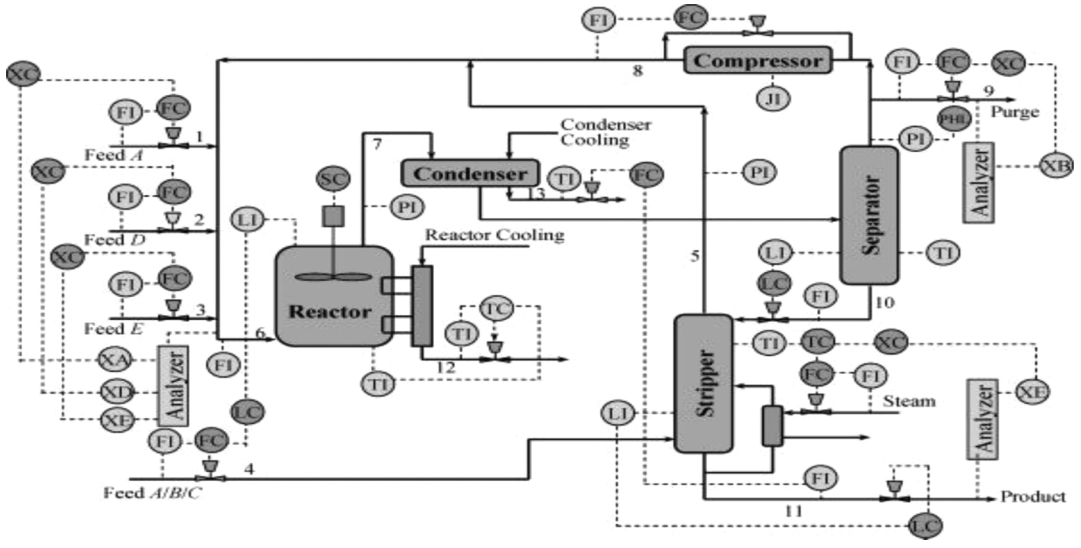

Figure 14. TE example: Flowchart of the TE process.

variation among variables. In fact, the dependence between x_1 and x_2 or between x_2 and x_3 has greatly changed.

Referring to online monitoring, a searching strategy is performed with the focus on the PDF values of the current monitored sample $\mathbf{X}_t^{\text{monitor}}$. For example, according to Table 3 and Figure 5, if $f^{(1)}(\mathbf{X}_t^{\text{monitor}}) = 0.8$, $f^{(2)}(\mathbf{X}_t^{\text{monitor}}) = 1.7$, and $f^{(3)}(\mathbf{X}_t^{\text{monitor}}) = 0.01$, then the GLP index corresponding to each mode satisfies $P_L^{(1)}(\mathbf{X}_t^{\text{monitor}}) \in (0.85, 0.9)$, $P_L^{(2)}(\mathbf{X}_t^{\text{monitor}}) \in (0.1, 0.15)$, and $P_L^{(3)}(\mathbf{X}_t^{\text{monitor}}) \in (0.9, 0.95)$, indicating that the current monitored sample $\mathbf{X}_t^{\text{monitor}}$ is closer to mode 2. Meanwhile, to obtain a certain value of the GBIP index rather than its estimated interval, GLP indices with regard to three modes are further set as the average of their upper and lower bounds, i.e., $P_L^{(1)}(\mathbf{X}_t^{\text{monitor}}) = 0.875$, $P_L^{(2)}(\mathbf{X}_t^{\text{monitor}}) = 0.125$, and $P_L^{(3)}(\mathbf{X}_t^{\text{monitor}}) = 0.925$.

The basic idea of the proposed VCDD monitoring approach is somewhat similar to that of the FGMM approach. Because both methods are aimed at modeling the joint distribution of

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 5. TE Example: Information on 22 Continuous Process Measurements

variables	process measurements	unit
XMEAS(1)	A feed (stream 1)	kscmh
XMEAS(2)	D feed (stream 2)	kg/h
XMEAS(3)	E feed (stream 3)	kg/h
XMEAS(4)	A and C feed (stream 4)	kscmh
XMEAS(5)	recycle flow (stream 8)	kscmh
XMEAS(6)	reactor feed rate (stream 6)	kscmh
XMEAS(7)	reactor pressure	kPa gauge
XMEAS(8)	reactor level	%
XMEAS(9)	reactor temperature	°C
XMEAS(10)	purge rate (stream 9)	kscmh
XMEAS(11)	product separator temp	°C
XMEAS(12)	product separator level	%
XMEAS(13)	product separator pressure	kPa gauge
XMEAS(14)	product separator underflow	m^3/h
XMEAS(15)	stripper level	°C
XMEAS(16)	stripper pressure	kPa gauge
XMEAS(17)	stripper underflow (stream 11)	m^3/h
XMEAS(18)	stripper temp	°C
XMEAS(19)	stripper steam flow	kg/h
XMEAS(20)	compress work	kW
XMEAS(21)	reactor cooling water outlet temp	°C
XMEAS(22)	separator cooling water outlet temp	°C

Table 6. TE Example: Information on Two Operating Modes

operating modes	G/H mass radio	product rate (stream 11)
1	50/50	G: 7038 kg/h, H: 7038 kg/h
3	90/10	G: 10000 kg/h, H: 1111 kg/h

each mode and then using the Bayesian inference technique to estimate the global probability-based monitoring index. However, the VCDD approach is very good at modeling the non-Gaussian mode (the deterministic relationship among the variables can also be highly nonlinear), and, more to the point, the DQA enables such a probability-based monitoring index to be estimated in real time. Note that the Mahalanobis distance or the corresponding local probability index of the FGMM approach defined in ref 35 is invalid when handling a specific non-Gaussian mode. In other words, from the perspective of the monitoring strategies, the VCDD approach that can deal with much more complex processes may be treated as an extended or generalized form of the FGMM approach. Therefore, the comparison results of these two approaches can provide compelling evidence on the advantages of the VCDD approach. Besides, different from the FGMM approach, the PCA-based method is also commonly used in process monitoring. Herein, the multiple PCA (MPCA) approach is used for comparison, and the T^2 and SPE statistics of the monitored sample are calculated within one certain mode after performing a clustering method, e.g., the Expectation-Maximization (EM) algorithm.

Figure 12 depicts the monitoring control charts of three approaches mentioned above. Herein, three evaluation indices, missed detection rate (MDR), false detection rate (FDR), and delay/detection time (DT), are given in Table 4. Two conventions are made as follows: (1) the DT of Fault 2 is estimated at the start of the 401st sample. If there are 5 consecutive monitoring index values exceeding the control limit, it is recognized as a fault that has been successfully triggered; (2) the FDR is calculated based on the overall 300 normal samples, from sample 1 to sample 50 and sample 151 to sample 400. That is to say, the FDR under each mode is not considered here.

The results show that the proposed C-VCDD approach achieves zero MDR, zero DT, and, at the same time, the lowest FDR among the three methods. The FDR of MPC (SPE statistics) (0.0433) is remarkably lower than that of FGMM (0.0933) but higher than that of C-VCDD (0.0267). Note that two faults studied here mainly reflect significant variation in residual space, hence, the T^2 statistics of the MPC approach are somewhat invalid.

Actually, the good monitoring performance of the C-VCDD approach is due to its trustworthy model constructed by C-vine copula. Figure 13 depicts the essence of the monitoring results. Referring to Fault 2 triggered in time period 401, the FGMM approach employs one Gaussian component to describe each mode, and the estimated 95% probability ellipse covers the training data in mode 1 with an additional blank area left, as shown in the top left subplot of Figure 13. This will obviously cause more faulty samples to be determined as normal ones and result in type II errors for the monitored data in that region. Yet the C-vine model correctly models the non-Gaussian process by describing the intrinsically complex dependence and has much more ability in detecting those faults reflecting significant shifts of variable dependence (though the amplitude of each variable does not change significantly). This is also the reason why the C-VCDD approach achieves lower FDR than FGMM. Since the bias of error on x_1 is large enough to keep away from the blank area, Fault 1 caused in time periods 51−150 can be completely detected by both approaches.

The MPCA approach decomposes data information into two subspaces, principle component subspace and residual subspace. The T^2 and SPE statistics are essentially the measurement of the Mahalanobis distance in each subspace, respectively. In ref 44, Yue and Qin pointed out that the PCA-based method is able to make a distinction between noise in the data and the latent states; in other words, this technique gives more tolerance and flexibility to the variation of noise in the residual subspace. Therefore, the PCA-based method is considered to be a robust monitoring technique. This directly explains why MPCA (SPE statistics) achieves lower FDR than FGMM. However, the PCA-based method still has limited ability in dealing with complex processes (correctly or incorrectly treats the non-Gaussian information as noise); note that the proposed VCDD approach has significantly lower FDR than MPCA. Overall, on the premise that the noise of training data is negligible, a trustworthy training model will play a much more important role in achieving a good monitoring performance than just separately considering different subspaces with less data information preserved.

5. APPLICATION TO THE TE PROCESS

In the process monitoring area, the TE process is one of the commonly used benchmarks. It was created by Eastman Chemical Company and can provide a realistic industrial process.[45] There are a reactor, a condenser, a compressor, a separator, and a stripper in this process. According to material balance, energy balance, and vapor−liquid equilibrium, differential equations are established for each unit. The flowchart of the process is shown in Figure 14.

There are 22 continuous process measurements, 19 composition measurements, 12 manipulated variables, and 21

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 7. TE Example: The Whole Optimized Pair Copulas in a C-Vine Model for Mode 1, the Copula Order of the 22 Variables Has Been Determined, Number 0 Denotes Independence Copula, the Rest of the Highlighted Numbers Represent Different Kinds of Bivariate Copulas

	x_{11}	x_{18}	x_{22}	x_{20}	x_{21}	x_{13}	x_7	x_{16}	x_9	x_5	x_6	x_{10}	x_{14}	x_{12}	x_{19}	x_8	x_{15}	x_4	x_{17}	x_1	x_2	x_3
x_{11}	\	1	1	10	1	1	1	1	14	5	13	1	5	5	0	0	0	0	0	0	0	0
x_{18}	\	\	0	4	0	1	1	1	5	5	1	2	1	0	0	0	0	0	0	1	0	0
x_{22}	\	\	\	1	10	5	0	0	0	14	0	0	0	0	0	0	0	0	0	0	0	0
x_{20}	\	\	\	\	33	1	1	1	0	0	13	5	0	0	0	0	0	0	0	0	0	0
x_{21}	\	\	\	\	\	5	5	5	1	0	0	4	0	0	0	0	0	0	0	40	5	1
x_{13}	\	\	\	\	\	\	1	1	0	13	0	10	0	0	5	0	0	0	0	0	23	0
x_7	\	\	\	\	\	\	\	1	0	0	0	0	0	0	0	0	0	0	0	0	0	5
x_{16}	\	\	\	\	\	\	\	\	0	0	0	5	0	0	0	0	0	0	0	5	0	0
x_9	\	\	\	\	\	\	\	\	\	0	0	0	0	0	0	0	0	0	1	14	0	33
x_5	\	\	\	\	\	\	\	\	\	\	0	0	0	0	0	0	0	0	0	0	0	0
x_6	\	\	\	\	\	\	\	\	\	\	\	0	5	0	0	0	0	0	0	0	0	0
x_{10}	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0	0	0	5	33	1	0	0
x_{14}	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0	0	0	0	0	0	0
x_{12}	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0	0	0	0	1	0
x_{19}	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0	0	0	0	0
x_8	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0	0	0	0
x_{15}	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0	0	0
x_4	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	1	0	0
x_{17}	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0
x_1	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0
x_2	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	1
x_3	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\

preprogrammed faults in the TE process. Moreover, corresponding to different G/H mass ratios in stream 11, there are six modes of TE process operating conditions. Note that the original process is strictly open-loop unstable, and the decentralized control design proposed by Ricker[46,47] is adopted for the closed-loop process operation. The sampling time is 0.05 h, and the 22 continuous process measurements are chosen as our monitored variables under mode 1 and mode 3. Detailed information on the 22 continuous process measurements and the parameters of mode 1 and mode 3 are given in Table 5 and Table 6, respectively. It is emphasized that two modes here are just randomly selected, and the proposed method can be successfully applied to the other four modes of the TE process as well.

When modeling, we first obtain 1000 normal data in mode 3 and switch the process to mode 1 and collect another 1000 data. Then we use C-vine copula to model these normal data. As for a C-vine model with 22 variables, 231 pair copulas should be considered. Since it is convenient to handle one pair copula at a time, the dependence description procedure actually appears not that cumbersome for high dimensional variables. Table 7 and Table 8 specify the chosen bivariate copulas for the two modes according to 2000 samples. See the simulation results of mode 1 in Table 7, the variable order has been determined, and the process variable x_{11} is set as the root node in tree 1. It is necessary to emphasize that though the most significantly correlated pairwise variables are x_7 and x_{16} whose Kendall rank correlation coefficient is 0.713, neither of them is selected as the root node due to eq 18. The related scatter plot is displayed in Figure 15(a), and it is found that there exists highly positive correlation with the reactor pressure and the stripper pressure. The numbers in Table 7 represent the family members (copula order) selected for the pair copulas, where number 0 stands for the independence copula and the remaining highlighted numbers represent different bivariate copulas optimized. For example, in tree 1, the first pair copula $c_{11,18}$ is chosen from 1 (Gaussian copula), the third pair copula $c_{11,20}$ is chosen from 10 (Frank-Joe/BB8 copula), the eighth pair copula $c_{11,9}$ is chosen from 14 (Survival Gumbel copula), the ninth pair copula $c_{11,5}$ is chosen from 5 (Frank copula), and the 10th pair copula $c_{11,6}$ is chosen from 13 (Survival Clayton copula); in tree 2, the 23rd pair copula $c_{18,20|11}$ is chosen from 4 (Gumbel copula), etc. More detailed information on the optimized bivariate copulas are given in the Supporting Information. It is noticed that among the 231 pair copulas, only 61 pair copulas in mode 1 and 53 pair copulas in mode 3 need to be optimized, while the rest of the pair copulas are fitted as independent ones whose copula density is $c = 1$. In essence, complex dependencies among high dimensional variables are decomposed into analyzing a series of bivariate copulas, and, more to the point, this flexible dependence structure is easy to handle since we can always solve the optimization problems in an analogous sparse matrix. Joint PDF values of the 1000 training data in each mode are

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Table 8. TE Example: The Whole Optimized Pair Copulas in a C-Vine Model for Mode 3

	x_{13}	x_7	x_{16}	x_{20}	x_{18}	x_{10}	x_{11}	x_{21}	x_{22}	x_5	x_6	x_{19}	x_2	x_{14}	x_3	x_1	x_9	x_{15}	x_{17}	x_4	x_{12}	x_8
x_{13}	\	2	2	24	2	1	34	1	1	5	1	0	0	0	0	0	0	0	0	0	0	0
x_7	\	\	1	13	1	0	1	13	0	0	0	0	0	0	0	0	0	0	0	0	0	0
x_{16}	\	\	\	19	5	0	5	5	30	23	1	0	0	0	0	1	0	0	0	0	0	0
x_{20}	\	\	\	\	0	0	23	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0
x_{18}	\	\	\	\	\	0	9	33	6	0	0	0	0	33	1	20	0	0	13	0	1	0
x_{10}	\	\	\	\	\	\	5	10	1	0	0	0	0	0	0	40	0	0	0	0	0	0
x_{11}	\	\	\	\	\	\	\	0	40	0	0	0	0	0	0	1	33	0	0	5	0	0
x_{21}	\	\	\	\	\	\	\	\	0	0	0	5	23	0	4	0	5	0	0	0	0	0
x_{22}	\	\	\	\	\	\	\	\	\	0	0	0	0	0	0	0	0	0	0	0	0	0
x_5	\	\	\	\	\	\	\	\	\	\	0	0	0	0	0	0	0	0	0	0	0	0
x_6	\	\	\	\	\	\	\	\	\	\	\	0	0	5	0	0	0	0	0	0	0	0
x_{19}	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0	0	0	0	0	0	0	0
x_2	\	\	\	\	\	\	\	\	\	\	\	\	\	0	1	0	0	0	0	5	0	
x_{14}	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	5	0	0	0	0	0	0
x_3	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0	0	0	0	0
x_1	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	34	0	5	0	0
x_9	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	1	0	0	0
x_{15}	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0	0
x_{17}	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0	0
x_4	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0	0
x_{12}	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	0
x_8	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\	\

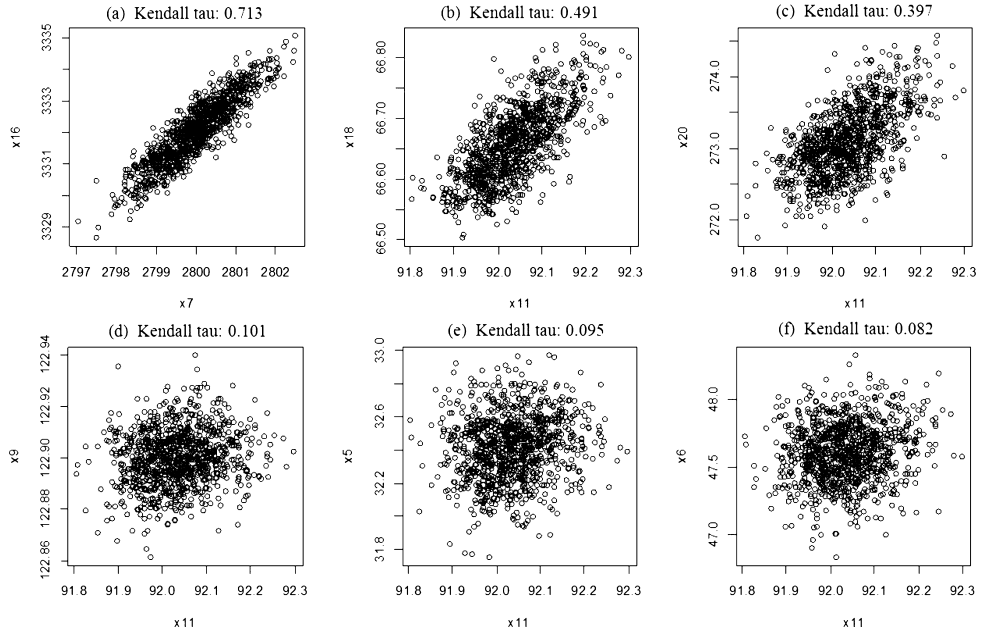

Figure 15. TE example: Scatter plots of some typical process variables, along with their Kendall rank correlation coefficients.

Table 9. TE Example: Information on Two Process Faults

variable	description	type
IDV (6)	a feed loss (stream 1)	step
IDV (13)	reaction kinetics	slow drift

calculated to obtain the density quantile table. The sampling process (e.g., the MCMC method) can be eliminated, thus vastly reducing the computation complexity.

Once the density quantile table has been created, the real-time monitoring begins. The monitored data contain 400 samples,

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

the first 100 samples come from mode 3 with normal condition, the 101st to 200th samples come from mode 3 with Fault 13, and the latter 200 samples are in mode 1 where Fault 6 is added in time period 301 to 400. Details about Fault 6 and Fault 13 are given in Table 9. Fault 13 is caused by the slow drift change of reaction kinetics, and Fault 6 is caused by the step change of A feed loss in stream 1. The GBIP control chart is shown in Figure 16. For the

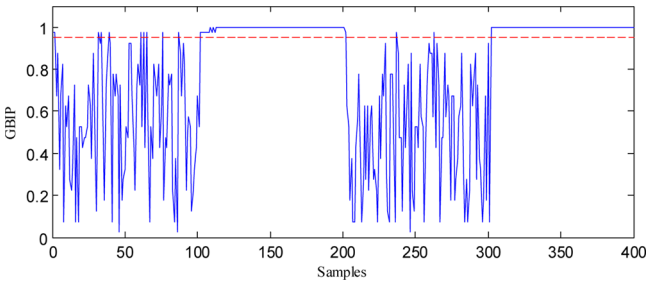

Figure 16. TE example: GBIP control chart with 95% control limit based on the C-VCDD approach.

normal data, the C-VCDD approach can hold a relatively lower false detection rate; when it comes to the faulty data, it can hold nearly 100% detection accuracy, indicating that the proposed C-VCDD approach performs well in TE process monitoring.

To analyze the computation complexity of the proposed C-VCDD approach, the CPU time deriving from three main aspects is analyzed: (1) variable order determination for each mode; (2) pair copula optimization for each mode; and (3) joint PDF value calculation for each mode. Herein, simulations are run on a 2.4 GHz CPU with 8 GB RAM computer using R (CDVine software package). As shown in Figure 17, it takes the most time in the process of pair copula optimization, i.e. 5.54 min for mode 1 and 5.11 min for mode 3. The vine copula optimization process for the 22 dimensional variables appears much more efficient and pragmatic than the traditional multivariate copulas that require solving a multivariate optimization problem in extremely high dimensions at a time. In addition, less than 2 min are needed to calculate the joint PDF values for the whole 2000 training data, indicating that the density quantile table is conveniently created. Overall, totally less than 15 min are needed in C-vine copula modeling for the two modes with the whole 2000 training data. Meanwhile, it is also found that the CPU time cost in calculating one monitored sample along with implementing the searching strategy in a brief table is approximately 0.1 s on average. Overall, in order to obtain the precise model of each mode, though the proposed C-VCDD approach needs a bit more time in offline modeling than the PCA-based and FGMM approach. The time cost in the online monitoring process is comparable with the other two approaches, which will make the C-VCDD monitoring approach a powerful and practical one.

6. CONCLUSIONS

In this article, a novel C-vine copula-based multimode process monitoring approach is proposed. In contrast to the commonly used strategy of dimensionality reduction, the proposed approach aims to create the statistical model of each mode by directly describing the complex dependencies among the process variables with high dimensionality, nonlinearity and non-Gaussian properties. Referring to online monitoring, a generalized local probability index is defined to measure the distance of the monitored data from each non-Gaussian mode. Meanwhile, a density quantile table is created, ensuring that the global GBIP index of the monitored data can be estimated by just searching a static table created offline.

The proposed C-VCDD approach is successfully applied in a numerical example and the TE benchmark process. Compared with MPCA and FGMM approaches, it is shown that the presented approach achieves remarkably lower false detection rates, missed detection rates, and delay time, especially for detecting one kind of drifting fault reflecting dependence variation among variables. It is also demonstrated that both offline modeling and online monitoring can achieve low computation load, which will make this approach more practical.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.iecr.5b01267.

Tables S1−S4 (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 86-21-64253820. E-mail: lishaojun@ecust.edu.cn.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The material is based upon work supported by the National Natural Science Foundation of China (under project No. 21176072) and the Fundamental Research Funds for the

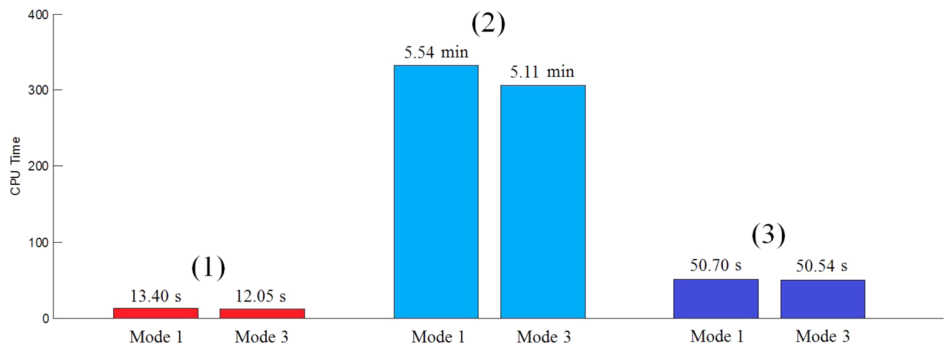

Figure 17. TE example: CPU time cost in (1) variable order determination, (2) pair copula estimation, and (3) joint PDF value calculation for mode 1 and mode 3.

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

Central Universities. The authors would like to thank Warren D. Seider and Ian H. Moskowitz for their invaluable comments on this work.

REFERENCES

(1) Venkatasubramanian, V.; Rengaswamy, R.; Yin, K.; Kavuri, S. N. A Review of Process Fault and Diagnosis, Part I: Quantitative Model-Based Methods. *Comput. Chem. Eng.* **2003**, *27*, 293−311.

(2) Venkatasubramanian, V.; Rengaswamy, R.; Kavuri, S. N. A Review of Process Fault and Diagnosis, Part II: Qualitative models and Search Strategies. *Comput. Chem. Eng.* **2003**, *27*, 313−326.

(3) Venkatasubramanian, V.; Rengaswamy, R.; Kavuri, S. N.; Yin, K. A Review of Process Fault and Diagnosis, Part III: Process History Based-Methods. *Comput. Chem. Eng.* **2003**, *27*, 327−346.

(4) Qin, S. J. Survey on Data-Driven Industrial Process Monitoring and Diagnosis. *Annu. Rev. Control.* **2012**, *36*, 220−234.

(5) Ge, Z. Q.; Song, Z. H.; Gao, F. R. Review of Recent Research on Data-Based Process Monitoring. *Ind. Eng. Chem. Res.* **2013**, *52* (10), 3543−3562.

(6) Wold, S.; Esbensen, K.; Geladi, P. Principal Component Analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2* (1), 37−52.

(7) MacGregor, J. F.; Jaeckle, C.; Kiparissides, C.; Koutoudi, M. Process Monitoring and Diagnosis by Multiblock PLS Methods. *AIChE J.* **1994**, *40*, 826−838.

(8) Lee, J. M.; Yoo, C. K.; Choi, S. W.; Vanrolleghem, P. A.; Lee, I. B. Nonlinear Process Monitoring Using Kernel Principal Component Analysis. *Chem. Eng. Sci.* **2004**, *59*, 223−234.

(9) Dong, D.; McAvoy, T. J. Nonlinear Principal Component Analysis-Based on Principal Curves and Neural Networks. *Comput. Chem. Eng.* **1996**, *20* (1), 65−78.

(10) Roweis, S. T.; Saul, L. K. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* **2000**, *290*, 2323−2326.

(11) Kano, M.; Tanaka, S.; Hasebe, S.; Hashimoto, I.; Ohno, H. Monitoring Independent Components for Fault Detection. *AIChE J.* **2003**, *49* (4), 969−976.

(12) Choi, S. W.; Park, J. H.; Lee, I. B. Process Monitoring Using a Gaussian Mixture Model via Principal Component Analysis and Discriminant Analysis. *Comput. Chem. Eng.* **2004**, *28* (8), 1377−1387.

(13) Ge, Z. Q.; Xie, L.; Song, Z. H. A Novel Statistical-Based Monitoring Approach for Complex Multivariate Processes. *Ind. Eng. Chem. Res.* **2009**, *48* (10), 4892−4898.

(14) Ge, Z. Q.; Song, Z. H. Process Monitoring Based on Independent Component Analysis-Principal Component Analysis (ICA-PCA) and Similarity Factors. *Ind. Eng. Chem. Res.* **2007**, *46* (7), 2054−2063.

(15) Zhang, Y. W. Fault Detection and Diagnosis of Nonlinear Processes Using Improved Kernel Independent Component Analysis (KICA) and Support Vector Machine (SVM). *Ind. Eng. Chem. Res.* **2008**, *47* (18), 6961−6971.

(16) Ge, Z. Q.; Yang, C. J.; Song, Z. H. Improved Kernel PCA-Based Monitoring Approach for Nonlinear Processes. *Chem. Eng. Sci.* **2009**, *64*, 2245−2255.

(17) Tian, Y.; Du, W. L.; Qian, F. Fault Detection and Diagnosis for Non-Gaussian Processes with Periodic Disturbance Based on AMRA-ICA. *Ind. Eng. Chem. Res.* **2013**, *52* (34), 12082−12107.

(18) Li, N.; Yan, W. W.; Yang, Y. P. Spatial-Statistical Local Approach for Improved Manifold-Based Process Monitoring. *Ind. Eng. Chem. Res.* **2015**, *54*, 8509.

(19) Escobar, M. S.; Kaneko, H.; Funatsu, K. Combined generative topographical mapping and graph theory unsupervised approach for nonlinear fault identification. *AIChE J.* **2015**, *61* (5), 1559−1571.

(20) Patton, A. J. A Review of Copula Models for Economic Time Series. *J. Multivariate Anal.* **2012**, *110*, 4−18.

(21) Patton, A. J. Copula Methods for Forecasting Multivariate Time Series. In *Handbook of Economic Forecasting*; Elsevier: Oxford, 2011.

(22) Weiß, G. N. F.; Supper, H. Forecasting Liquidity-Adjusted Intraday Value-At-Risk with Vine Copulas. *J. Bank Financ.* **2013**, *37* (9), 3334−3350.

(23) Cherubini, U.; Luciano, E.; Vecchiato, W. *Copula Methods in Finance*; Wiley: England, 2004.

(24) Scholzel, C.; Friederichs, P. Multivariate Non-Normally Distributed Random Variables in Climate Research - Introduction to The Vopula Approach, 2008. Available at https://hal-cea.archives-ouvertes.fr/cea-00440431/ (accessed April 3, 2015).

(25) Aas, K.; Berg, D. Models for Construction of Multivariate Dependence - A Comparison Study. *Eur. J. Financ.* **2009**, *15* (7−8), 639−659.

(26) Meel, A.; Seider, W. D. Plant-Specific Dynamic Failure Assessment Using Bayesian Theory. *Chem. Eng. Sci.* **2006**, *61*, 7036−7056.

(27) Pariyani, A.; Seider, W. D.; Oktem, U. G.; Soroush, M. Dynamic Risk Analysis Using Alarming Databases to Improve Process Safety and Product Quality: Part I-Data Compaction. *AIChE J.* **2012**, *58* (3), 812−825.

(28) Pariyani, A.; Seider, W. D.; Oktem, U. G.; Soroush, M. Dynamic Risk Analysis Using Alarming Databases to Improve Process Safety and Product Quality: Part II-Bayesian Analysis. *AIChE J.* **2012**, *58* (3), 826−841.

(29) Ahooyi, T. M.; Soroush, M.; Arbogast, J. E.; Seider, W. D.; Oktem, U. G. Maximum-Likelihood Maximum-Entropy Constrained Probability Density Function Estimation for Prediction of Rare Events. *AIChE J.* **2014**, *60* (3), 1013−1026.

(30) Nelson, R. B. *An Introduction to Copulas*, 2nd ed.; Springer: Berlin, 2006.

(31) Joe, H. Families of m-variate Distributions with Given Margins and m(m-1)/2 Bivariate Dependence Parameters. In *Distributions with Fixed Marginals and Related Topics*; Ruschendorf, L., Schweizer, B., Taylor, M. D., Eds.; Institute of Mathematical Statistic: Hayward, 1996; pp 120−141.

(32) Bedford, T.; Cooke, R. M. Vines - A New Graphical Model for Dependent Random Variables. *Ann. Stat.* **2002**, *30*, 1031−1068.

(33) Kurowicka, D.; Cooke, R. M. *Uncertainty Analysis with High Dimensional Dependence Modeling*; John Wiley & Sons: Chichester, 2006.

(34) Aas, K.; Czado, C.; Feigessi, A.; Bakken, H. Pair-Copula Construction of Multiple Dependence. *Insurance Math. Econom.* **2009**, *44* (2), 182−198.

(35) Yu, J.; Qin, S. J. Multimode Process Monitoring with Bayesian Inference-Based Finite Gaussian Mixture Models. *AIChE J.* **2008**, *54* (7), 1811−1829.

(36) Hyndman, R. J. Computing and Graphing Highest Density Regions. *Am. Stat.* **1996**, *50* (2), 120−126.

(37) Sklar, A. Fonctions dé Repartition á *n* Dimension et Leurs Marges. *Publ. Inst. Stat. Univ. Paris* **1959**, *8*, 229−231.

(38) Kurowicka, D.; Joe, H. *Dependence Modeling Vine Copula Handbook*; World Scientific: Singapore, 2011.

(39) Schirmacher, D.; Schirmacher, E. Multivariate Dependence Modeling Using Pair-Copulas, 2008. Available at http://scholar.g363.com/scholar?q=Multivariate+Dependence+Modeling+Using+Pair-Copulas (accessed April 3, 2015).

(40) Ahooyi, T. M.; Arbogast, J. E.; Soroush, M. Rolling Pin Method: Efficient General Method of Joint Probability Modeling. *Ind. Eng. Chem. Res.* **2014**, *53* (52), 20191−20203.

(41) Brechmann, E. C.; Schepsmeier, U. Modeling Dependence with C- and D- Vine Copulas: The R Package CDVine. *J. Stat. Softw.* **2013**, *52* (3), 1−27.

(42) Schepsmeier, U.; Brechmann, E. C. Statistical Inference of C- and D-vine Copulas, 2014. Available at http://cran.r-project.org/ (accessed April 3, 2015).

(43) Ren, X.; Li, S. J.; Lv, C.; Zhang, Z. Y. Sequential Dependence Modeling Using Bayesian Theory and D-Vine Copula and Its Application on Chemical Process Risk Prediction. *Ind. Eng. Chem. Res.* **2014**, *53* (38), 14788−14801.

(44) Yue, H. H.; Qin, S. J. Reconstruction-Based Fault Identification Using A Combined Index. *Ind. Eng. Chem. Res.* **2001**, *40* (20), 4403−4414.

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX

(45) Downs, J. J.; Vogel, E. F. A Plant-Wide Industrial Process Control Problem. *Comput. Chem. Eng.* **1993**, *17* (3), 245−255.
(46) Ricker, N. L. Decentralized Control of The Tennessee Eastman Challenge Process. *J. Process Control* **1996**, *6* (4), 205−221.
(47) Ricker, N. L. Tennessee Eastman Challenge Archive, 2002. Available at http://depts.washington.edu/control/LARRY/TE/download.html (accessed April 3, 2015).

DOI: 10.1021/acs.iecr.5b01267
Ind. Eng. Chem. Res. XXXX, XXX, XXX−XXX