doi:10.11835/j.issn.1000-582X.2021.124

基于 LASSO 回归的 R-vine copula 模型构建及其在化工过程故障检测中的应用

邓红涛[1],贾 琼[2],李绍军[2],李 伟[1]

(1. 石河子大学,新疆 石河子 832000;
2.华东理工大学 化工过程先进控制和优化技术教育部重点实验室,上海 200237)

摘要:Vine copula 模型在描述高维数据间的非线性、非高斯特性相依关系问题上提供了一种新的思路,在化工过程建模领域受到越来越多关注。笔者将 LASSO(least absolute shrinkage and selection operator)回归引入 R-vine copula(LASSO-R-vine copula,LRVC),根据变量间联系的强弱程度确定变量在 R-vine 矩阵中的位置,利用回归分析正则化路径选择 R-vine copula 矩阵结构,遵循 R-vine 矩阵构建规则和回归过程确定 R-vine 结构矩阵模型,以获得一个与变量独立性有关的稀疏矩阵模型。该方法构建的矩阵结构独立于 copula 函数类型和参数,在处理高维度复杂工业过程数据时,利用稀疏模型和惩罚力度简化 copula 函数类型选择过程,缩短了建模时间,使统计建模具有更强的灵活性。该方法在 TE(Tennessee Eastman)和醋酸脱水过程故障监测中表现出较好的预测效果,证明了提出的方法在非线性、非高斯过程的有效性。

关键词:过程监控;相关性;R-vine copula;LASSO 回归

中图分类号:TP277 **文献标志码**:A **文章编号**:1000-582X(2023)01-027-08

Model of R-vine copula based on LASSO regression and its application in chemical process fault detection

DENG Hongtao[1], *JIA Qiong*[2], *LI Shaojun*[2], *LI Wei*[1]

(1. Shihezi University, Shihezi, Xinjiang 832000, P. R. China; 2. Key Laboratory of Advanced Control and Optimization for Chemical Processes Under the Ministry of Education, East China University of Science and Technology, Shanghai 200237, P. R. China)

Abstract: The vine copula model provides a new way to describe the nonlinear and non-Gaussian dependence of high-dimensional data and has attracted more and more attention in the field of chemical process modeling. In this article, a novel chemical process fault detection method, LASSO-R-vine copula (LRVC), is proposed by introducing LASSO (least absolute shrinkage and selection operator) regression into R-vine copula. LRVC determines the position of the variables in the R-vine matrix according to the strength of the relationship between the variables, using regression to analyze the regularization path and select the R-vine copula matrix structure. The R-vine structure matrix model is determined to obtain a sparse matrix model

收稿日期:2021-04-20 **网络出版日期**:2021-09-16

基金项目:国家自然科学基金资助项目(21676086)。
Supported by the National Natural Science of China (21676086).

作者简介:邓红涛(1981—),女,硕士,副教授,主要从事数据处理、过程建模、优化与控制研究,(E-mail)38190425@qq.com。

通信作者:李伟,男,硕士,副教授,主要从事数据处理、智能控制方向研究,(E-mail)lw200@qq.com.cn。

related to variables' independence by following the R-vine matrix construction rules and regression process. The matrix structure constructed by this method is independent of the copula function type and parameters. When dealing with high-dimensional complex industrial process data, sparse models and penalties could simplify the copula function type's selection process, shorten the modeling time, and make the statistical modeling more flexible. This method shows an excellent predictive effect in TE and the acetic acid dehydration process fault monitoring, proving its effectiveness in nonlinear and non-Gaussian processes.

Keywords: process monitoring; correlation; R-vine copula; LASSO regression

过程安全和产品质量是目前化工过程被关注最多的两个问题,而过程监控是提高过程安全和产品质量的重要手段。随着集散控制和数据采集系统的广泛应用,工业过程采集数据的维度和数据量不断增加,导致基于经验知识和解析模型的方法在工业过程监控领域的研究受到了限制[1]。基于数据驱动的故障监控建模方法通过统计与分析过程数据来挖掘系统的特性,在描述未知机理模型和缺乏过程知识的复杂过程问题研究中备受学术和工业界关注。传统的数据分析方法有偏最小二乘法(PLS)[2]和主元分析法(PCA)[3],这两种方法适合分析具有线性、高斯分布特性数据。而处理非线性数据时常使用核方法来对这两种方法进行改进,进而形成核偏最小二乘(KPLS)法[4]、核主元分析(KPCA)法[5]等。这些方法主要是基于降维的方式将高维数据映射到低维特征空间,消除变量间的相关性,但在降维过程中数据携带信息量都会有一定的损失。

近年来,针对复杂数据的相关性研究中,copula 理论得到了广泛的应用,将联合概率分布函数与边缘概率分布函数之间的相关性结构建立联系。由于多变量 copula 在构建维数较大的数据间的依赖性时缺乏灵活性,Joe[6]提出了 pair-copula,该方法是将多变量 copula 用一系列的二元 copula 来表示,该方法在刻画高维数据的条件相关性、非对称性、尾部相关性等方面均体现出更大的灵活性,已经在金融、环境、工程等领域得到了广泛的应用。2015 年 Ren 等[7]首先将 vine copula 函数引入到化工过程监控领域,提出了基于 vine copula 相关性描述的多模态故障检测方法。Zheng 等[8]利用 D-vine copulas 混合模型实现了对复杂的工业过程监控。周南等[9]利用核密度估计法构建 R-vine copula 结构并用在工业过程故障检测中。由于二元 copula 函数类型众多,vine copula 结构矩阵不固定,选取合适的 vine copula 结构矩阵和 copula 函数类型成为建模的关键环节。目前常规方法是利用贪婪算法计算所有可能矩阵结构下变量之间的相关关系,选择相关性最大的矩阵结构,然后根据赤池准则(AIC)选取 copula 函数类型和参数[10-12]。这种方法在处理高维的工业过程数据时,会出现计算量大、计算时间长的问题,其解也不能保证接近该高维组合优化问题的最优解。

笔者在构建 R-vine copula 模型时引入 LASSO 回归来统计变量之间的相关关系,根据变量间联系的强弱程度确定变量在 R-vine 矩阵中的位置,利用回归分析正则化路径选择 R-vine copula 矩阵结构。遵循 R-vine矩阵构建规则和回归过程中惩罚力度调整变量在矩阵中的位置,确定 R-vine 结构矩阵模型,以获得一个与变量独立性有关的稀疏矩阵模型。该方法构建的矩阵结构独立于 copula 函数类型和参数,在处理高维度复杂工业过程数据时,利用稀疏模型和惩罚力度简化 copula 函数类型选择过程,缩短了建模时间,使统计建模具有更强的灵活性。该方法应用在 TE(Tennessee Eastman)过程中表现出较好的检测效果。

1 R-vine copula 理论基础和 LASSO 回归

1959 年 Sklar 第一次提出用 copula 函数分析复杂变量间的相关关系,将多维变量的联合分布函数用边缘分布函数和 copula 函数表示,但是这种 copula 函数面对高维数据时会出现参数过多而导致优化困难的问题[11]。1996 年 Joe[6]提出了 vine copula 结构。Vine copula 结构分解具有较大的灵活性,分解策略较多[13-14]。2002 年 Bedford 等[15]定义了 R-vine copula 结构分解模型,相应的多元分布结构称为 R-vine 结构,可以更加灵活地表达变量之间的关系[16]。

1.1　构建 R-vine 结构矩阵规则

用树结构来表示 vine 结构，对于 n 维变量的 R-vine 分布，包含了 $n-1$ 棵树，每棵树由节点和边组成。两个节点确定一条边，每条边用二元 copula 函数表示。由于 vine copula 结构分解模型多样，树型结构不唯一。为了更好地表示 vine copula 的结构，2013 年 Brechmann 等[17]研究了一种 R-vine copula 结构模型，利用一个下三角矩阵 $\boldsymbol{M}$ 来表示 R-vine 分解模型，用矩阵 $\boldsymbol{M}$ 可以简单地表示出 R-vine 结构中的树集 T、节点集 V 和边集 E。

对一个 n 维变量的 vine copula 结构，可以用一个 $n\times n$ 的下三角矩阵 $\boldsymbol{M}$ 表示，矩阵对角线元素 m 代表变量 $X=\{x_1,x_2,\cdots,x_n\}$ 中的 x_m。用 $m_{i,j}$ 表示第 i 行第 j 列矩阵元素，$m_{i,j}\in\{1,2,\cdots,n\}$，$i=(1,2,\cdots,n)$，$j=(1,2,\cdots,i)$。矩阵元素之间需要满足以下条件：

1）$\{m_{j,j},\cdots,m_{n,j}\}\subset\{m_{i,i},\cdots,m_{n,i}\}$，$1\leqslant i<j\leqslant n$。即第 j 列元素集$\{m_{j,j},\cdots,m_{n,j}\}$属于第 i 列元素集$\{m_{i,i},\cdots,m_{n,i}\}$。

2）$m_{j,j}\notin\{m_{i,i},\cdots,m_{n,i}\}$，$1\leqslant j<i\leqslant n$。即第 i 列元素集$\{m_{i,i},\cdots,m_{n,i}\}$不包含第 j 列对角线上元素 $m_{j,j}$。

3）$\{m_{i,j},\{m_{i+1,j},\cdots,m_{n,j}\}\}\subseteq\{m_{k,k},\{m_{w,k},\cdots,m_{n,k}\}\}$。即如果矩阵元素集$Q=\{m_{i,j},\{m_{i+1,j},\cdots,m_{n,j}\}\}$，$j=(1,2\cdots,n-2)$，$i=(j+1,\cdots,n)$，那么一定存在包含或等于 Q 的元素集$\{m_{k,k},\{m_{w,k},\cdots,m_{n,k}\}\}$，$j<k$，$w\geqslant i$。

设满足以上规则的矩阵 $\boldsymbol{M}$ 中第 i 行第 j 列元素 m 集合为 W。

$$\boldsymbol{M}=\begin{pmatrix}5&&&&\\2&\bar{4}&&&\\1&2&\bar{3}&&\\③&\bar{1}&\bar{2}&\bar{2}&\\4&\bar{3}&\bar{1}&\bar{1}&1\end{pmatrix}$$

图 1　5 维变量的 R-vine 矩阵 $\boldsymbol{M}$

Fig. 1　R-vine matrix $\boldsymbol{M}$ of 5-dimensional variables

根据以上条件可以发现矩阵中元素位置是互相约束的，如图 1 所示矩阵 $\boldsymbol{M}$，组成第 2 列的变量$\{4,2,1,3\}$包含在第 1 列$\{5,2,1,3,4\}$中；第 2 列对角线元素$\{4\}$不会出现在第 3、4、5 列中；第 1 列画圈元素组成元素集 $Q=\{3,4\}$，那么至少第 2 列中存在包含 Q 的元素集$\{4,1,3\}$。

假设画圈元素 $m_{4,1}$ 为 O，O 和第 1 列中行数大于 4 的元素组成元素集 $Q=\{O,4\}$。列数大于 1、行数大于等于 4 的元素和对角线元素（带上横线元素）组成的元素个数大于等于 Q 的元素集$\{4,1,3\}$、$\{3,2,1\}$、$\{2,1\}$，这些元素集中包含元素 4 的只有$\{4,1,3\}$，根据条件 3 可知画圈位置的元素 O 只能在变量集 $W=\{3,1\}$中选取。

1.2　利用矩阵 $\boldsymbol{M}$ 表示联合密度函数

利用 R-vine 的矩阵结构，不需要画树结构就可以快捷地表示出多维变量的联合密度函数[18]。矩阵元素 $m_{i,j}$ 和第 j 列元素可以表示树 T_{n-i+1} 的第 $i-j+1$ 个二元 copula 函数 $c_{j(e),k(e)\mid D(e)}(F_{j\mid D}(x_j(e)\mid x_{D(e)}),F_{k\mid D}(x_{k(e)}\mid x_{D(e)}))$，其中 $j(e)=m_{j,j}$，$k(e)=m_{i,j}$，$D(e)=\{m_{i+1,j},\cdots,m_{n,j}\}$，$i=(2,3,\cdots,n)$，$j=(1,2,\cdots,i-1)$，$F$ 为条件累计分布函数。

n 维变量的联合概率密度 P 可以表示为边缘概率密度 $f(x_i)$ 与 copula 密度函数 c 的乘积形式：

$$P=f(x_1,x_2,\cdots,x_n)=\prod_{i=1}^{n}f_i(x_i)\times\prod_{i=2}^{n}\prod_{j=1}^{i-1}c_{m_{j,j},m_{i,j}\mid m_{i+1,j},\cdots,m_{n,j}}(F(x_{m_{j,j}}\mid x_{m_{i+1,j},\cdots,m_{n,j}}),F(x_{m_{i,j}}\mid x_{m_{i+1,j},\cdots,m_{n,j}}))。\tag{1}$$

式中：$f_i(x_i)$ 为第 i 个变量 x_i 的边缘概率密度函数，$c_{m_{j,j},m_{i,j}\mid m_{i+1,j},\cdots,m_{n,j}}$ 为二元 copula 函数，$F(x_{m_{j,j}}\mid x_{m_{i+1,j};m_{n,j}})$ 为 $x_{m_{j,j}}$ 的条件累计分布函数。

图 1 所示矩阵 $\boldsymbol{M}$ 为 R-vine 结构矩阵，矩阵中元素 m 的值 1－5 代表变量 $X=\{X_1,X_2,X_3,X_4,X_5\}$。矩阵 M 中画圈元素 $m_{4,1}=3$，对应二元 copula 为：$c_{3,5\mid 4}$，变量 $m_{1,1}$ 条件密度可以用第 1 列所有元素的二元 copula 函数之积和对角线元素的边缘密度表示：

$$f(x_5\mid x_1x_2x_3x_4)=f(x_5)c_{4,5}c_{3,5\mid 4}c_{1,5\mid 3,4}c_{2,5\mid 1,3,4}。\tag{2}$$

联合概率密度为：

$$f(x_1,x_2,x_3,x_4,x_5)=f(x_1)f(x_2\mid x_1)\times f(x_3\mid x_1x_2)f(x_4\mid x_1x_2x_3)f(x_5\mid x_1x_2x_3x_4)。\tag{3}$$

根据矩阵 $\boldsymbol{M}$ 求 5 维变量联合概率密度，即：

$$f(x_1,x_2,x_3,x_4,x_5)=f(x_1)f(x_2)f(x_3)f(x_4)f(x_5)\times c_{1,2}c_{1,3}c_{3,4}c_{4,5}c_{2,3|1}c_{1,4|3}c_{3,5|4}c_{2,4|1,3}c_{1,5|3,4}c_{2,5|1,3,4}。\tag{4}$$

在满足矩阵规则的前提下构造合理的R-vine结构矩阵 $\boldsymbol{M}$ 是构建R-vine copula模型的关键[19]。目前常用贪婪算法构建R-vine树的结构矩阵,计算树节点之间的相关性,按照最强相依性原则遍历所有可能的矩阵结构,寻找最大相关系数之和以构建R-vine结构矩阵,并在整个结构选择过程中迭代拟合copula函数及其参数。此方法计算时间长,计算量大,构建的模型结构复杂,容易出现过拟合现象。

1.3 LASSO回归

线性回归表示变量间相互依赖的定量关系,变量 $x=\{x_1,x_2,\cdots,x_n\}$ 回归函数如下:

$$h=\sum_{m=1}^{n}\varphi m x_m=\varphi_1x_1+\varphi_2x_2+\cdots+\varphi_nx_n,\tag{5}$$

式中 φ_m 为变量 x_m 的回归系数。

LASSO回归是一种数据降维方法,善于处理变量的筛选。1996年Tibshirani[20]首次在普通线性回归模型中添加了惩罚项,通过改变惩罚项将一些作用比较小的变量线性系数压缩,最终变为零,从而获得稀疏解。这种基于惩罚方法对样本数据进行变量选择,防止了数据过拟合,不但可以用于线性关系,也可以用于非线性关系。回归损失函数公式如下:

$$J(\varphi)=\frac{1}{2q}\sum_{t=1}^{q}(h_\varphi(x^{(t)})-y^{(t)})^2+\lambda\sum_{k=1}^{r}|\varphi k|,\tag{6}$$

式中:$h_\varphi(x^{(t)})$ 是根据式(5)计算预测第 t 个样本的值,因变量 y 为真实样本值,q 为样本个数,$t=(1,2,\cdots,q)$,λ 为正则化参数,r 为参数个数,$k=(1,2,\cdots,r)$。随着 λ 增大,各变量的系数逐渐趋于零。

2 利用LASSO回归构建R-vine矩阵

将LASSO回归算法用于构建R-vine结构矩阵,提出一种新的构建R-vine结构矩阵的方法[21]。如构建 n 维变量 $X=\{X_1,X_2,\cdots,X_n\}$ 的R-vine结构矩阵 $\boldsymbol{M}$。

首先确定R-vine结构矩阵 $\boldsymbol{M}$ 中的对角线元素。将变量带入式(7),按照式(6)利用交叉验证方法确定过程变量 $x_i(i=1,2,\cdots,n)$ 的LASSO回归方程。

$$\begin{cases}x_1=\varphi_{1,2}x_2+\varphi_{1,3}x_3+\cdots+\varphi_{1,z}x_z+\cdots+\varphi_{1,n}x_n,\\ x_2=\varphi_{2,1}x_1+\varphi_{2,3}x_3+\cdots+\varphi_{2,z}x_z+\cdots+\varphi_{2,n}x_n,\\ \vdots\\ x_m=\varphi_{m,1}x_1+\varphi_{m,3}x_3+\cdots+\varphi_{m,z}x_z+\cdots+\varphi_{m,n}x_n,\\ \vdots\\ x_n=\varphi_{n,1}x_1+\varphi_{n,2}x_2+\cdots+\varphi_{n,z}x_z+\cdots+\varphi_{n,n-1}x_{n-1}。\end{cases}\tag{7}$$

式中:$m=1,2,\cdots,n$,$z=1,2,\cdots,n$,且 $m\neq z$。根据式(8)统计 n 个回归方程中变量 x_m 的回归系数 $\varphi_{m,z}$ 非零的个数 u_z,按照 u_z 的升序将变量 x_m 设为R-vine结构矩阵 $\boldsymbol{M}$ 的对角线元素,当出现次数相同时按照回归系数的和排序。

$$u_z=\sum_{m=1}^{n}1_{\{\varphi_{m,z}\neq0,m\neq z\}}。\tag{8}$$

如图1所示,矩阵 $\boldsymbol{M}$ 确定对角线元素时,存在 $u_5\leqslant u_4\leqslant u_3\leqslant u_2\leqslant u_1$,那么对角线元素为{5,4,3,2,1}。

接下来按照从右往左、从下至上的顺序确定矩阵中的其他元素。如确定矩阵 $\boldsymbol{M}$ 中的变量 $m_{i,j}$,以矩阵对角线元素 $x_{m_{j,j}}$ 为因变量 y,根据R-vine结构矩阵构建规则,以满足元素集 W 的元素为自变量,带入式(5),根据式(9)利用最小回归角方法选择变量,当回归系数依次置为零时,惩罚项最大的变量即为 $m_{i,j}$。

$$\min_{\varphi}\left(\frac{1}{2q}\sum_{t=1}^{q}(x_{m_{i,j}}^{(t)}-\sum_{l\in W}\varphi_l x_l^{(t)})^2+\sum_{l\in W}\lambda_l|\varphi_l|\right),\tag{9}$$

式中:W 为根据R-vine矩阵规则确定的此处可放置的变量集合,l 是属于集合 W 的元素。

确定图1中矩阵 $\boldsymbol{M}$ 第5行第4列的元素,根据R-vine矩阵规则可知 $m_{5,4}=1$,然后按照从右往左、从下

至上的顺序依次确定 $m_{5,3}$、$m_{5,2}$、$m_{5,1}$、$m_{4,3}$、$m_{4,2}$、$m_{4,1}$、$m_{3,2}$、$m_{3,1}$、$m_{2,1}$。确定画圈元素 $m_{4,1}$ 时，根据构建 R-vine结构矩阵规则推出可用变量 $W=\{1,3\}$，则以 $m_{1,1}$ 为自变量 y，以 $m_{4,2}$ 和 $m_{5,5}$ 为因变量，按照式(5)利用 LASSO 回归计算回归系数，按照式(9)利用最小回归角方法调整惩罚项。当 $\varphi_1=0$ 时 $\lambda=\lambda_1$；$\varphi_3=0$ 时 $\lambda=\lambda_3$。因为存在 $\lambda_1<\lambda_3$，那么 $m_{4,1}=3$。

3　基于 LASSO 构建 R-vine copula 模型的故障检测方法

本研究中利用 LASSO 回归过程反映变量之间关系的特点，按照变量回归系数归零速度和惩罚项大小确定 R-vine 结构矩阵 $\boldsymbol{M}$，利用正常样本数据根据赤池准则确定 R-vine copula 模型中参数，构建模型，利用阈值法进行在线故障检测(如图 2 所示)。

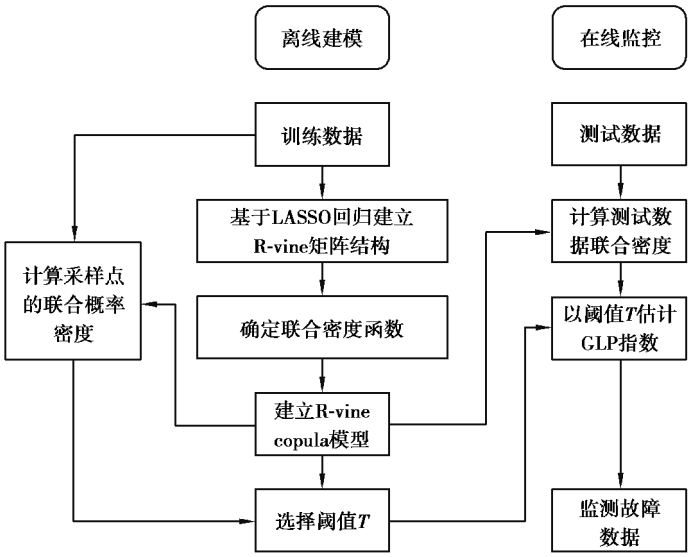


图 2　LRVC 过程监控方法示意图

Fig. 2　Flow chart of the LRVC monitoring approach

3.1　离线建模

1)获得正常操作过程的训练样本集，按照第 2 节方法构建 R-vine 矩阵 $\boldsymbol{M}$。

2)根据 R-vine 矩阵 $\boldsymbol{M}$ 确定 copula 对和参数，构建 R-vine copula 模型。合适的 copula 对能够精确地描述变量数据间的相关关系。采用基于似然函数的赤池信息准则[23]选取最合适 copula 对类型。赤池准则是权衡被估计模型复杂度和拟合优越性的一种标准，其定义如下：

$$\mathrm{AIC}=-2\log L(\hat{\theta})+2\eta。\tag{10}$$

式中：η 是 copula 函数的参数个数，一般为 1～2 个；L 为似然函数；$\hat{\theta}$为 copula 函数参数的估计值。根据赤池准则选最小 copula 类型确定 R-vine copula 模型。根据式(1)计算样本数据联合概率密度。

3)确定检测阈值 T。

计算样本点的联合概率密度，利用分位数法[7]求对应的联合概率密度中阈值 T。该方法根据高密度区域与密度分位数理论构建广义贝叶斯推断概率指标(GLP)，阈值 T 选取 99%的控制限，监测时对比静态密度分位数表确定监测状态。

3.2　在线监控

1)利用模型计算监测数据联合概率密度函数。

2)以阈值 T 为界限，小于阈值则为故障数据。

4 应用分析

4.1 TE过程

Eastman公司依照实际的化工反应过程开发了TE测试平台，仿真数据具有非线性、时变和强耦合性等特征，能很好地模拟复杂工业过程，被广泛应用于控制、优化、过程监控与故障诊断的研究。TE数据集由训练集和测试集构成，数据集包含了正常状态数据和21种故障状态数据，每个样本都有52个观测变量，其中连续变量22个，操纵变量12个和成分变量19个。本研究所用数据可从http://web.mit.edu/braatzgroup/links.html下载，在离线状态下选取了正常工况下样本52个变量中的22个连续的过程变量构建R-vine copula模型，采用960个正常样本点来建立模型，每种故障状态的测试数据也选用960个样本点进行测试，对21个故障数据进行监测。将本研究中提出的方法LRVC与独立成分分析(ICA)、高斯混合模型(FGMM)、R-vine copula(RVC)等算法计算TE过程故障检测率进行比较，结果见表1，其中T^2和SPE是ICA故障检测的统计指标，T^2指标衡量样本向量在主元空间的变化，SPE指标衡量样本向量在残差空间的投影的变化，BIP是贝叶斯推理的后验概率(BIP)指标，FT指利用RVC模型和LRVC模型下的广义贝叶斯推断概率指标。表中粗体表示检测效果最好的值。

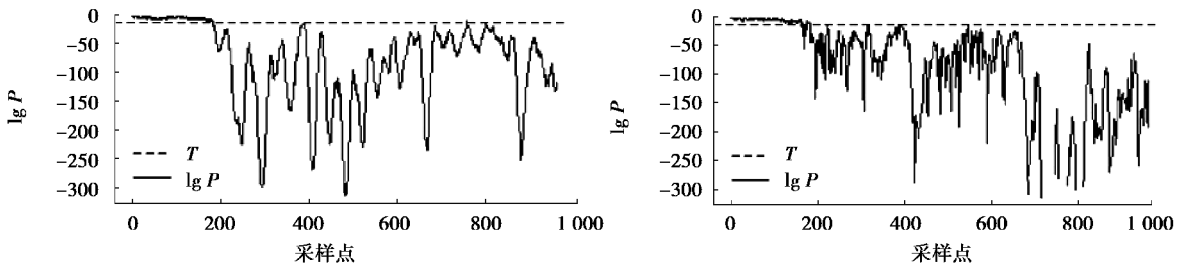
表1 TE过程故障检测率对比表

Table 1 Fault detection rates of TE process with different monitoring approaches

故障序号	ICA		FGMM	RVC	LRVC	故障序号	ICA		FGMM	RVC	LRVC
	T^2	SPE	BIP	FT	FT		T^2	SPE	BIP	FT	FT
1	99.62	**99.88**	**99.88**	99.50	99.63	12	96.75	99.12	98.50	98.38	**98.75**
2	97.75	98.12	98.43	98.63	**100**	13	94.25	**94.75**	94.63	94.38	94.38
3	1.50	**6.50**	1.76	0.25	6.25	14	99.88	**100**	99.88	99.88	99.88
4	1.13	3.13	2.23	0.38	7.75	15	1.38	5.37	**8.30**	0.25	2.13
5	24.37	26.50	24.50	22.25	**26.75**	16	20.37	32.37	21.75	16.50	**23.38**
6	**100**	**100**	**100**	**100**	**100**	17	**93.75**	91.63	93.37	86.75	89.25
7	34.13	36.63	35.78	36.63	**40.63**	18	89.75	**90.63**	89.75	89.25	89.38
8	96.25	**97.75**	97.62	97.00	**97.75**	19	**35.13**	15.50	10.88	6.50	18.00
9	1.38	5.37	3.71	0.25	**6.13**	20	77.00	67.50	70.25	75.50	**80.13**
10	72.75	70.75	70.87	48.13	70.00	21	19.63	**58.13**	38.25	40.50	44.13
11	27.63	**31.87**	19.75	19.63	28.40						

可以看出利用LASSO回归构建的R-vine模型监测结果优于贪婪算法构建R-vine矩阵建模，与FGMM、ICA方法[23-24]比较具有较高的检测率。LRVC模型的故障检测率都略高于利用贪婪算法构建矩阵的模型检测率，在故障2、4～9、12、16、20中都表现出较好的检测效果，其他故障的检测效果和其他方法相差不多。LRVC模型有13类故障的检测率都高于FGMM方法的检测结果，相比于ICA检测方法在故障2、4、5、7、9、12、20中都表现较好检测率。

TE过程中数据具有非高斯态的特性，FGMM方法是基于马氏距离判断数据是否异常，反映数据的非高斯特性能力较差；ICA方法在数据变换和特征提取的过程中会造成部分信息的损失。而LRVC方法利用LASSO回归建立R-vine矩阵构建模型，全面挖掘出数据变量之间的信息，在刻画非高斯、非线性方面有显著的优势，提高了故障检测的性能。图3为LRVC方法对故障8和故障12的监控图，横坐标为对联合概率密度P取对数。



(a)故障8　(b)故障12

图 3　TE 过程故障 8、12 的 LRVC 监控图

Fig. 3　Monitoring diagrams of TE process failures 8 and 12

4.2　醋酸脱水过程

醋酸脱水过程控制系统包括 4 个进料和 90 级塔板，用于建模的 300 组测试数据和 500 组训练数据来自分布式控制系统(DCS)，包括温度、流量、压力等连续的 21 个监测变量。表 2 为核主元分析 KPCA、FGMM、和 LRVC 3 种方法的检测率(fault detection rate，FDR)和误报率(false positive rate，FPR)，可见 LRVC 方法在醋酸脱水过程监测中表现较好。

表 2　醋酸脱水过程检测率和误报率

Table 2　Fault detection and false alarm rates of the acetic acid dehydration process

指标	KPCA		FGMM	LRVC
	T^2	SPE	GLP	GLP
FDR	1	0.990	1	1
FPR	0.11	0.045	0.18	0.03

5　结　论

提出了一种基于 LASSO 回归构建 R-vine copula 模型的化工过程故障检测方法 LRVC，与传统方法相比具有较好的检测效果。LRVC 模型利用 LASSO 回归过程反映变量之间相关性的特性，依据回归过程惩罚力度构建 R-vine 矩阵，采用赤池准则进行 copula 类型选择构建 R-vine copula 模型。模型中利用 LASSO 回归统计多元变量相关性构建 R-vine 矩阵更能体现 vine copula 结构分解特性，适合高维变量数据分析。基于 LRVC 模型的过程故障检测方法充分利用了 copula 函数可以反映变量之间非高斯、非线性的特性，在不降维的情况下直接描述变量之间关系，具有更好的解释能力和适应性。此方法简化了构建 R-vine 矩阵过程，在超高维变量的建模过程中可以直接利用矩阵进行 copula 函数的选择，节省了建模时间。该方法在 TE 过程以及醋酸脱水过程中的应用表明其在工业过程故障检测中具有应用前景。

参考文献：

[1] 曹立立. 基于 HMM 的 TE 过程在线故障诊断与多步故障预报[D]. 武汉：华中科技大学，2015.
Cao L L. HMM-based on-line fault diagnosis and multi-step ahead fault prediction for TE process[D]. Wuhan: Huazhong University of Science & Technology, 2015. (in Chinese)

[2] MacGregor J F, Jaeckle C, Kiparissides C, et al. Process monitoring and diagnosis by multiblock PLS methods[J]. AIChE Journal, 1994, 40(5): 826-838.

[3] Kramer M A. Nonlinear principal component analysis using autoassociative neural networks[J]. AIChE Journal, 1991, 37(2): 233-243.

[4] Zhang Y W, Hu Z Y. Multivariate process monitoring and analysis based on multi-scale KPLS[J]. Chemical Engineering Research and Design, 2011, 89(12): 2667-2678.

[5] Lee J M, Yoo C, Choi S W, et al. Nonlinear process monitoring using kernel principal component analysis[J]. Chemical Engineering Science, 2004, 59(1): 223-234.

[6] Joe H. Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters[J]. Lecture Notes-Monograph Series, 1996, 28: 120-141.

[7] Ren X, Tian Y, Li S J. Vine copula-based dependence description for multivariate multimode process monitoring[J]. Industrial & Engineering Chemistry Research, 2015, 54(41): 10001-10019.

[8] Zheng W J, Ren X, Zhou N, et al. Mixture of D-vine copulas for chemical process monitoring[J]. Chemometrics and Intelligent Laboratory Systems, 2017, 169: 19-34.

[9] 周南, 李绍军. 基于核密度估计的 R-Vine Copula 选择及其在故障检测中的应用[J]. 高校化学工程学报, 2019, 33(2): 443-452.
Zhou N, Li S J. R-Vine Copula selection based on kernel density estimation and its application in fault detection[J]. Journal of Chemical Engineering of Chinese Universities, 2019, 33(2): 443-452.(in Chinese)

[10] Haff H I, Aas K, Frigessi A, et al. Structure learning in Bayesian Networks using regular vines[J]. Computational Statistics & Data Analysis, 2016, 101: 186-208.

[11] Genest C, Favre A C. Everything you always wanted to know about copula modeling but were afraid to ask[J]. Journal of Hydrologic Engineering, 2007, 12(4): 347-368.

[12] Kurowicka D, Joe H. Dependence modeling: vine copula handbook[M]. New Jersey: World Scientific, 2011.

[13] Hyvärinen A, Oja E. Independent component analysis: algorithms and applications[J]. Neural Networks, 2000, 13(4/5): 411-430.

[14] Cossette H, Gadoury S P, Marceau É, et al. Hierarchical Archimedean copulas through multivariate compound distributions[J]. Insurance: Mathematics and Economics, 2017, 76: 1-13.

[15] Bedford T, Cooke R M. Probability density decomposition for conditionally dependent random variables modeled by vines [J]. Annals of Mathematics and Artificial Intelligence, 2001, 32(1/2/3/4): 245-268.

[16] Chang B, Joe H. Prediction based on conditional distributions of vine copulas[J]. Computational Statistics & Data Analysis, 2019, 139: 45-63.

[17] Brechmann E C, Schepsmeier U. Modeling dependence with C- and D-vine copulas: the R package CDVine[J]. Journal of Statistical Software, 2013, 52(3): 1-27.

[18] Dißmann J, Brechmann E C, Czado C, et al. Selecting and estimating regular vine copulae and application to financial returns[J]. Computational Statistics & Data Analysis, 2013, 59: 52-69.

[19] Müller D, Czado C. Dependence modelling in ultra high dimensions with vine copulas and the Graphical Lasso[J]. Computational Statistics & Data Analysis, 2019, 137: 211-232.

[20] Tibshirani R. Regression shrinkage and selection via the lasso[J]. Journal of the Royal Statistical Society: Series B (Methodological), 1996, 58(1): 267-288.

[21] Müller D, Czado C. Election of sparse vine copulas in high dimensions with the Lasso[J]. Computational Statistics and Data Analysis, 2019, 29: 269-287.

[22] Akaike H. Information theory and an extension of the maximum likelihood principle[M]// Selected papers of Hirotugu Akaike. New York: Springer, 1998: 199-213.

[23] Lee J M, Yoo C K, Lee I B. Statistical process monitoring with independent component analysis[J]. Journal of Process Control, 2004, 14(5): 467-485.

[24] Yu J, Qin S J. Multimode process monitoring with Bayesian inference-based finite Gaussian mixture models[J]. AIChE Journal, 2008, 54(7): 1811-1829.

(编辑 罗 敏)