Computers and Chemical Engineering 115 (2018) 141–149

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Quality-relevant independent component regression model for virtual sensing application

Xinmin Zhang[a], Manabu Kano[a,*], Yuan Li[b]

[a] Department of Systems Science, Kyoto University, Kyoto 606-8501, Japan
[b] Department of Information Engineering, Shenyang University of Chemical Technology, Shenyang 110142, PR China

ARTICLE INFO

Article history:
Received 26 October 2017
Revised 21 March 2018
Accepted 3 April 2018
Available online 6 April 2018

Keywords:
Virtual sensing
Independent component regression
Distance covariance and distance correlation
Quality-relevant independent component regression
Vinyl acetate monomer process
Near-infrared spectral analysis

ABSTRACT

Independent component regression (ICR) is an efficient method for tackling non-Gaussian problems. In this work, the defects of the conventional ICR are analyzed, and a novel quality-relevant independent component regression (QR-ICR) method based on distance covariance and distance correlation is proposed. QR-ICR extracts independent components (ICs) using a quality-relevant independent component analysis (QR-ICA) algorithm, which simultaneously maximizes the non-Gaussianity of ICs and statistical dependency between ICs and quality variables. Meanwhile, two new types of statistical criteria, called cumulative percent relevance (CPR) and Max-Dependency (Max-Dep), are proposed to rank the order and determine the number of ICs according to their contributions to quality variables. The proposed QR-ICR$_{(CPR)}$ and QR-ICR$_{(Max\text{-}Dep)}$ methods were validated through a vinyl acetate monomer production process and a benchmark near-infrared spectral data. The results have demonstrated that the proposed QR-ICR$_{(CPR)}$ and QR-ICR$_{(Max\text{-}Dep)}$ provide simpler predictive models and give better prediction performances than PLS, ICR, ICR$_{(CPR)}$, and ICR$_{(Max\text{-}Dep)}$.

1. Introduction

In any industry, it is important to monitor some important quality variables in real time. However, in some cases, online measurement of quality variables is difficult or even impossible due to limitations of hardware analyzers, which are time-consuming maintenance, needing for calibration, aged deterioration, long dead-time, and slow dynamics as clarified via the questionnaire survey in Japan (Kano and Fujiwara, 2013). To solve these problems, virtual sensing technology has been developed and widely used in industrial processes. The basic idea of virtual sensing technology is to estimate the product quality utilizing some easily measured secondary variables. In recent years, data-driven soft-sensors have attracted much more attention because a massive amount of process data are recorded. Among them, multiple linear regression (MLR), principal component regression (PCR), and partial least squares (PLS) are the most popular multivariate statistical modeling approaches (Kadlec et al., 2009; Kano and Ogawa, 2010).

In complex chemical processes, the conventional methods like PCR and PLS are inadequate to describe the underlying data structure because they are second-order methods, which consider only mean and variance or covariance of process data, and cannot cope with non-Gaussianity. To deal with this problem, some non-Gaussian methods like support vector machine (Balabin and Lomakina, 2011; Xu et al., 2017), Gaussian mixture model (Yu and Qin, 2008), and independent component analysis (ICA) (Hyvärinen, 1999) have been developed. Among them, ICA is an effective technique for extracting several independent signal components from observed data (Hyvärinen and Oja, 2000; Oja and Nordhausen, 2001). Compared to principal component analysis (PCA), ICA gives a high-order representation for non-Gaussian data. The recovered independent components (ICs) reveal more useful statistical information and make the interpretation of regression model easier as compared to principal components (PCs). ICA has been used for process monitoring, and it is demonstrated that ICA-based methods are superior to PCA-based methods (Kano et al., 2004; 2003; Lee et al., 2004). For quality control, ICA-based regression methods such as independent component regression (ICR) have been proposed. ICR uses the extracted ICs as predictor variables to build a regression model based on the least square criterion. Chen and Wang (2001) used ICR to separate the constituent components from the mixed spectra and estimate their concentrations. Westad (2005) combined ICA and PLS (IC-PLS) for sensory data

* Corresponding author.
E-mail addresses: zhang.xinmin.58s@st.kyoto-u.ac.jp (X. Zhang), manabu@human.sys.i.kyoto-u.ac.jp (M. Kano), li-yuan@mail.tsinghua.edu.cn (Y. Li).

https://doi.org/10.1016/j.compchemeng.2018.04.006
0098-1354/

analysis. Gustafsson (2005) illustrated that ICR yields chemically interpretable latent variables in multivariate regression as compared to PCR and PLS. Shao et al. (2006) have further demonstrated that ICR is a promising tool to retrieve both quantitative and qualitative information from the near-infrared spectra (NIR) data. Kaneko et al. (2008) applied an ICA-MLR method to aqueous solubility data analysis. Zhao et al. (2010) introduced an improved ICR and quantitative calibration algorithm through constructing joint latent variables based regression model. Ge and Song (2014) developed an ensemble ICR model based on different ensemble strategies. Zhang et al. (2015) proposed an ICR-based just-in-time learning model for complex batch process quality prediction. Zheng and Song (2016) proposed a two-level ICR model based on Bayesian inference strategy for multivariate spectroscopic calibration.

The conventional ICR also shows some shortcomings. For example, after extracting ICs from the observed data, there is no standard criterion to decide how many ICs should be retained for building an accurate prediction model. Unlike PCA which ranks the PCs in descending order according to the variance of each PC, ICA cannot decide which ICs are more significant because the extracted ICs have equal variances. In fact, to ensure optimal prediction performance, it is necessary for the engineers to know the magnitude of contribution of each IC to the regression model. In addition, uninformative ICs should not be used in regression models to improve their interpretability, prediction performance, and computational load. Furthermore, for regression analysis, the prediction performance of regression models like ICR actually depends on how well the features (ICs) are extracted from observed data. However, in the conventional ICR algorithm, the extracted ICs take no account of the quality information, thus it may not be optimal for quality prediction.

To deal with these issues, we propose a novel quality-relevant independent component regression (QR-ICR) method based on distance covariance and distance correlation. The basic idea is to first estimate ICs using a quality-relevant independent component analysis (QR-ICA) algorithm which simultaneously maximizes both the non-Gaussianity of ICs and statistical dependency between ICs and quality variables. Second, two new types of statistical criteria, called cumulative percent relevance (CPR) and Max-Dependency (Max-Dep), are designed to rank the order and determine the number of ICs. Compared to the conventional ICR method, QR-ICR can extract ICs which have more useful information on product quality, and thus it is beneficial to improve the prediction performance. In addition, by using CPR or Max-Dep criterion, QR-ICR (here denote as QR-ICR$_{(CPR)}$ or QR-ICR$_{(Max\text{-}Dep)}$) gives a much simpler calibration model with fewer significant ICs, and thus the interpretation of the model becomes easier, and the computational load for online prediction becomes lower.

The rest of this paper is organized as follows. Section 2 gives a brief introduction of distance covariance and distance correlation, and ICR. Then, the quality-relevant ICR (QR-ICR) algorithm with two new types of statistical criteria: cumulative percent relevance (CPR) and Max-Dependency (Max-Dep) is presented in Section 3. In Section 4, the proposed QR-ICR$_{(CPR)}$ and QR-ICR$_{(Max\text{-}Dep)}$ methods are applied to a vinyl acetate monomer production process and a benchmark near-infrared spectral data, and their application results are compared with PLS, ICR, ICR$_{(CPR)}$, and ICR$_{(Max\text{-}Dep)}$. The conclusions of this work are presented in Section 5.

2. Existing methods

2.1. Distance covariance and distance correlation

Classical Pearson product-moment covariance and correlation (ρ) measure linear dependence between two random variables, and in the bivariate normal case $\rho = 0$ is equivalent to independence. Distance covariance (dcov) and distance correlation (dcor) are new statistics that measure dependence of random vectors, without requiring linear normality assumptions (Székely et al., 2007; 2009). The idea of correlation is generalized in the following fundamental ways:

(1) $dcor(X, Y)$ is defined for X and Y in arbitrary dimensions.
(2) $dcor(X, Y) = 0$ if and only if the random vectors are independent.
(3) $dcor(X, Y)$ does not require distributional assumptions.
(4) In the bivariate normal case, $dcor(X, Y)$ is a deterministic function of (ρ), and $dcor(X, Y) \le |\rho|$ with equality when $\rho = \pm 1$.

Given two random vectors $X \in \Re^p$ and $Y \in \Re^q$, let $F_X(t)$ and $F_Y(s)$ denote the characteristic functions of X and Y, and $F_{X,Y}(t, s)$ be the joint characteristic function of X and Y. The population distance covariance between X and Y is the nonnegative number $dcov(X, Y)$ defined by

$$dcov^2(X,Y) = \int_{\Re^{p+q}} \|F_{X,Y}(t,s) - F_X(t)F_Y(s)\|^2 \omega(t,s)dtds \tag{1}$$

$$\omega(t,s) = (c_p c_q \|t\|_p^{1+p} \|s\|_q^{1+q})^{-1} \tag{2}$$

$$c_d = \pi^{(1+d)/2} / \Gamma\{(1+d)/2\}, \quad d \in \{p, q\} \tag{3}$$

where $\Gamma(\cdot)$ is the complete gamma function and $\|\cdot\|_d$ denotes the Euclidean norm of random vector in $\Re^d$. Similarly, the distance variances of X and Y are defined as

$$dcov^2(X) = \int_{\Re^{p+q}} \|F_{X,X}(t,s) - F_X(t)F_X(s)\|^2 \omega(t,s)dtds \tag{4}$$

$$dcov^2(Y) = \int_{\Re^{p+q}} \|F_{Y,Y}(t,s) - F_Y(t)F_Y(s)\|^2 \omega(t,s)dtds \tag{5}$$

The population distance correlation between X and Y with finite first moments is the nonnegative number $dcor(X, Y)$ defined by

$$dcor^2(X,Y) = \begin{cases} \dfrac{dcov^2(X,Y)}{\sqrt{dcov^2(X)dcov^2(Y)}}, & dcov^2(X)dcov^2(Y) > 0 \\ 0, & dcov^2(X)dcov^2(Y) = 0 \end{cases} \tag{6}$$

To calculate the distance covariance, a natural estimator of dcov on the basis of the distance dependence statistics, according to Székely et al. (2007, 2009), is given by

$$\widehat{dcov}^2(X,Y) = S_1 + S_2 - 2S_3 \tag{7}$$

$$S_1 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \|X_i - X_j\| \, \|Y_i - Y_j\| \tag{8}$$

$$S_2 = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \|X_i - X_j\| \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{N} \|Y_i - Y_j\| \tag{9}$$

$$S_3 = \frac{1}{N^3} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{l=1}^{N} \|X_i - X_l\| \, \|Y_j - Y_l\|. \tag{10}$$

where $\{(X_i, Y_i), i = 1, 2, \cdots, N\}$ denotes a random sample from the population (X, Y), and N is the number of sample vectors. Similarly, we can calculate the sample distance variance $\widehat{dcov}^2(X)$ and