Chemometrics and Intelligent Laboratory Systems 130 (2014) 115–122

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

# Ensemble independent component regression models and soft sensing application

Zhiqiang Ge, Zhihuan Song *

*State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, Department of Control Science and Engineering, Zhejiang University, Hangzhou 310027, Zhejiang, China*

ARTICLE INFO

*Article history:*
Received 4 May 2013
Received in revised form 7 September 2013
Accepted 19 September 2013
Available online 27 September 2013

*Keywords:*
Independent component regression
Ensemble learning model
Soft sensor

ABSTRACT

Independent component regression (ICR) model is able to extract underlying components, while simultaneously models high order statistics from the non-Gaussian process data. Based on different ensemble strategies, this paper aims to develop various ensemble forms of the ICR model. Specifically, by re-sampling of data samples, a bagging ICR model is developed; based on the independent component decomposition, a subspace ICR model is constructed through each direction of independent components; a further bi-dimensional ensemble ICR model is then constructed by combining bagging and subspace ICR models through two ensemble directions. For online measurements of key variables in industrial processes, various soft sensors are built based on different ensemble ICR models. Performance evaluations and detailed comparative studies of the developed soft sensors are illustrated through an industrial process.



## 1. Introduction

Due to limitations of the measurement technique and instrumentations, some key variables in the process industry are difficult to measure online, which are often determined by offline analyses in the laboratory or by some online analyzers. However, both offline analyses and online analyzers are expensive and time-consuming, which may introduce a large delay to the control system. Therefore, many soft sensing methods have been developed for estimation and prediction of those important variables by using other easy-to-measure process variables [1].

Principal component regression (PCR) and partial least squares (PLS) are two of the most widely used methods for soft sensing in industrial process. Even though successful studies have been reported on PCR and PLS based soft sensing methods, it has been pointed out that neither of these two methods can efficiently recover the underlying linear latent model for the process data [2]. Besides, PCR/PLS can only model the first and the second order statistics for the process data, which means the higher order statistical information has been ignored. However, the first and second order statistics can only describe the process data which is Gaussian distribution. For non-Gaussian process data, high order statistics are necessary for data description and modeling [3].

Independent component analysis (ICA) is a relatively new data analysis method, which is first proposed in the signal processing area [3]. The main idea of ICA is to recover true underlying sources from mixed signals, which are also statistically independent from each other. Different from the uncorrelated nature of the principal components extracted by PCR/PLS, independence is a much stronger condition, which can make better use of higher order statistical signals. Therefore, ICA is capable of modeling the process data which is non-Gaussian distributed [4–10]. Recently, ICA has been employed for regression purpose, which is known as independent component regression (ICR) [11–20].

However, it has been explored that a single calibration model does not always provide satisfactory predictions in practice. Instead, by combing the results of multiple regression models which are developed for the same purpose, the prediction accuracy of the calibration model could probably be improved. Actually, this is the idea of ensemble learning, which has recently caught much attention in the community of machine learning and pattern recognition. Conventional ensemble learning techniques include bagging, boosting, and the random subspace method [21–28]. The ensemble learning based method is especially applicable for unstable models, such as models which are sensitive to small changes of the data, initializations of the modeling procedure, etc. [29]. Due to the random initialization procedure, the independent components extracted by the ICA algorithm may not be reproducible, which will simultaneously cause unstable prediction results of the ICR model. Therefore, ensemble learning is potentially useful to improve the performance of the ICR model.

Among the three categories of the ensemble learning method, bagging and boosting models are both developed based on some types of re-sampling algorithms on the training data samples, while the random subspace model is constructed based on re-sampling of process variables. In the present work, bagging and random subspace based ensemble learning methods are incorporated into the ICR model to improve the prediction performance of the soft sensor. First, by re-sampling the training data samples for the ICR model, a bagging

* Corresponding author. Tel.: +86 571 87951442.
*E-mail address:* zhsong@iipc.zju.edu.cn (Z. Song).

0169-7439/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.chemolab.2013.09.009

ICR regression model is developed. Second, improved from the random subspace method, a new subspace ICR model is built for soft sensor modeling. Furthermore, a bi-dimensional ensemble ICR model is constructed, which is based on the re-sampling algorithm through both of the sample and variable directions. Detailed comparative studies of these ensemble forms of the ICR model are provided in Section 5 of this paper.

The remainder of this paper is structured as follows. In Section 2, a brief description of the ICR model for soft sensing is given. Three different ensemble learning ICR models are developed in Section 3, which is followed by soft sensing methods based on the developed ensemble ICR models. In Section 5, effectiveness of the ensemble ICR models is evaluated through an industrial case study. Finally, conclusions are made.

## 2. Independent component regression

Based on the ICA algorithm, it is assumed that the measured process variables $\mathbf{x} \in R^{m \times 1}$ can be expressed as linear combinations of $r(\leq m)$ unknown independent components $\mathbf{s} \in R^{r \times 1}$, which is [3]

$$\mathbf{x} = \mathbf{A}\mathbf{s} + \mathbf{e} \tag{1}$$

where $\mathbf{A} \in R^{m \times r}$ is the mixing matrix, $\mathbf{e} \in R^{m \times 1}$ is the residual. The aim of ICA is to estimate the original source $\mathbf{s}$ and the mixing matrix $\mathbf{A}$ from $\mathbf{x}$. Correspondingly, the objective of ICA is to calculate a separating matrix $\mathbf{W}$ so that the components of the reconstructed data matrix $\hat{\mathbf{s}}$ become as independent of each other as possible, which is given as

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x} \tag{2}$$

The number of independent components can be determined by various methods, such as non-Gaussianity test, variance based method, and so on [3]. When the original source $\mathbf{s}$ has been estimated from the process data, a linear regression model can be developed between the dataset $\hat{\mathbf{S}} = s[\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \cdots, \hat{\mathbf{s}}_n]^T \in R^{n\tilde{n}r}$ and the quality variable dataset $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]^T \in R^{n \times p}$, which can be calculated as

$$\mathbf{Q} = \left(\hat{\mathbf{S}}^T\hat{\mathbf{S}}\right)^{-1}\hat{\mathbf{S}}^T\mathbf{Y} \tag{3}$$

Denote the dataset of process variables as $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n]^T \in R^{n \times m}$, and combine the two steps of ICR modeling procedures, the ICR regression matrix can be calculated as

$$\mathbf{R}_{\mathrm{ICR}} = \mathbf{Q}^T\mathbf{W} \tag{4}$$

## 3. Ensemble independent component regression models

In this section, detailed descriptions of three different ensemble ICR models are given. First, a bagging form of the ICR model is proposed, which is developed through re-sampling of the sample direction. Second, a subspace ICR model is constructed, which is based on the re-sampling algorithm through the variable direction of the process data. Third, by re-sampling on both sample and variable directions, a bi-dimensional ensemble form of the ICR model is developed.

### 3.1. Bagging ICR model

Instead of making predictions from a single ICR model, the bagging ICR model constructs a series of ICR models for utilization. Each ICR model is developed from a re-sampled set of the original training dataset. In order to improve the accuracy and robustness of the prediction model, the prediction results from different ICR models are combined together. To illustrate the bagging ICR modeling procedures, denote the process dataset as $\mathbf{X} \in R^{n \times m}$, the quality dataset as $\mathbf{Y} \in R^{n \times p}$, where $n$ is the number of data samples, $m$ is the process variable number, and $p$ is the number of quality variables. Based on the re-sampling algorithm, data samples are randomly selected to form different re-sampled sets. Suppose the size of each re-sampled set is $n_b$ and the number of bagging ICR models is $B$, the re-sampled set can be represented as $\mathbf{X}_b = \mathbf{X}[rand(n_b)] \in R^{n_b\tilde{n}m}$, where $b = 1, 2, \cdots, B$, $rand(n_b)$ means randomly selecting $n_b$ data samples from $\mathbf{X}$. Then the ICR model can be constructed on the re-sampled dataset $\mathbf{X}_b$, which is given as

$$\begin{aligned} \mathbf{X}_b^T &= \mathbf{A}_b\mathbf{S}_b + \mathbf{E}_b \\ \mathbf{Q}_b &= \left(\mathbf{S}_b^T\mathbf{S}_b\right)^{-1}\mathbf{S}_b^T\mathbf{Y} \\ \mathbf{R}_{b,ICR} &= \mathbf{Q}_b^T\mathbf{W}_b \end{aligned} \tag{5}$$

where $b = 1, 2, \cdots, B$, $\mathbf{A}_b \in R^{m\tilde{n}r_b}$ and $\mathbf{W}_b \in R^{r_b\tilde{n}m}$ are mixing and separating matrices of the ICR model, $r_b$ is the selected number of independent components in each model, $\mathbf{R}_{b,\mathrm{ICR}}$ is the ICR regression matrix.

### 3.2. Subspace ICR model

Different from the bagging method, the random subspace based ensemble method carries out the re-sampling through the variable direction of the process data. Therefore, a series of subspaces which consist of different variables are constructed for ICR modeling. However, it has recently been noticed that the random subspace method has a serious drawback, which is, the random selection of process variables in each subspace does not guarantee the modeling performance of the subspace model. Actually, a good ensemble model should build individual subspace models which are not only accurate but also diverse. However, if the subspace is determined by randomly selecting process variables, different subspaces may have high overlap, thus the diversity of the subspace model cannot be guaranteed. In the present work, an improved subspace modeling strategy is proposed, which is based on the independent component decomposition of the process variables. In this case, different subspaces are then built through the independent component directions determined by the ICA algorithm, which are orthogonal to each other. Therefore, the diversity of the subspace model can be greatly improved. Furthermore, if the most important variables in each independent component subspace are selected, the accuracy of the subspace model can also be obtained. Detailed description of this improved subspace modeling strategy with the ICR model is given as follows.

Given the whole process dataset as $\mathbf{X} \in R^{n \times m}$, where $m$ is the number of process variables, and $n$ is the sample number for each variable. An initial independent component decomposition is carried out on $\mathbf{X}$, thus

$$\mathbf{X}^T = \mathbf{A}\mathbf{S} + \mathbf{E} \tag{6}$$

where $\mathbf{A} \in R^{m \times r}$ is the mixing matrix, $\mathbf{S} \in R^{r \times n}$ is the independent component (IC) matrix, and $\mathbf{E}$ is the residual matrix. Based on the orthogonal behavior of the extracted independent components, a subspace can be constructed through each IC direction. In order to select the dominant process variables in each subspace, the importance of each variable in different subspaces should be measured. To this end, an independent component related index (ICRI) is defined as follows

$$\mathbf{ICRI}(i,j) = \frac{a_{ij}^2}{a_{1j}^2 + a_{2j}^2 + \cdots + a_{ij}^2 + \cdots + a_{mj}^2} \tag{7}$$

where $i = 1, 2, \cdots, m$, $j = 1, 2, \cdots, r$, $a_{ij}$ is the $i$-th element of the $j$-th independent component direction in the mixing matrix $\mathbf{A}$. Based on the ICRI index, the importance of each process variable in different