Chemometrics and Intelligent Laboratory Systems 155 (2016) 160–169

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Two-level independent component regression model for multivariate spectroscopic calibration

Junhua Zheng, Zhihuan Song *

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

ARTICLE INFO

Article history:
Received 21 August 2015
Received in revised form 2 April 2016
Accepted 5 April 2016
Available online 12 April 2016

Keywords:
Multivariate calibration
Independent component regression
Two-level modeling
Ensemble learning
Bayesian inference

ABSTRACT

In this paper, a two-level independent component regression (ICR) model is developed for multivariate spectroscopic calibration. Compared to the traditionally used principal component regression and partial least squared regression model, the ICR model is more efficient to extract high order statistical information from the spectra data. To improve the calibration performance, an ensemble form of the ICR model is proposed. In the first level of the method, various subspaces are constructed based on the independent component decomposition of the original data space. Meanwhile, by defining a related index, the most important variables in each subspace are selected for ICR modeling, which form the second level of the proposed method. A Bayesian inference strategy is further developed for probabilistic combination of calibration results obtained from different subspaces. For performance evaluation, two case studies are carried out on a benchmark spectra dataset.

1. Introduction

In the past several decades, spectroscopic calibration modeling has become an effective tool for fast and non-invasive analysis in chemistry/biochemistry related areas, such as petrochemical and food industries and pharmaceutical and biological sectors [1–5]. In order to guarantee a high calibration performance for the spectroscopic device, various chemometrics methods have been incorporated. Commonly used ones include principal component regression (PCR), partial least squares (PLS), artificial neural networks (ANN), support vector regression (SVR), Gaussian process regression (GPR), etc. [6–20]. Among all those developed chemometrics modeling methods, the linear calibration model PCR and PLS may be two of the most widely used and accepted ones, which are efficient to provide fast and linear relationship analyses between the spectra and properties of the products.

Though successful studies have demonstrated the efficiency of PCR and PLS based calibration methods, as Gustafsson pointed out, neither of these two methods can generally recover a true underlying linear latent model from the data [21]. In addition, PCR/PLS can only extract the first and second order statistics from the data, which means higher order statistical information has been ignored. For non-Gaussian data, high order statistics are necessary for information extraction and interpretation. As an emergent data analysis technique in recent years, independent component analysis (ICA) aims to decompose the original signals into different directions, which are independent to each other [22]. The extracted component by the ICA model is assumed to be mutually independent instead of merely uncorrelated. Through probability interpretation, independence is a much stronger condition than uncorrelatedness, which can make use of higher order statistical information. Compared to PCR/PLS, it has been demonstrated that the ICA regression method (ICR) can recover the true underlying sources much better, depending on which an improved statistical interpretation of the data can be obtained. Recent works on ICA or ICA regression (ICR) have been done for blind source signal separation, image processing, process monitoring, spectra data analysis and quality prediction [23–33].

However, most ICR model based spectroscopic calibration works have been carried out on constructing a single model, no matter how complicated the spectra data performed. This may cause unsatisfactory performance, especially when the number of training data samples is relatively small, compared to the number of spectra data variables. Inspired by the idea of ensemble learning from the area of machine learning, the calibration performance could probably be improved through constructing multiple regression models for the same purpose. Typical ensemble learning strategies include bagging, random subspace, random forest, etc. [34–40]. In this paper, the random subspace method is employed and combined with ICR for model calibration purpose. The main idea of the random subspace method is to build various individual models based on different variable subsets which are randomly selected from the original variables space. Then, the final calibration result is obtained by combining the results of different individual models. However, a critical shortcoming of this method is that the diversity among different individual models cannot be well guaranteed through a random selection manner, which is a quite important issue in the ensemble learning method.

* Corresponding author.
E-mail addresses: jhzheng@zju.edu.cn (J. Zheng), zhsong@iipc.zju.edu.cn (Z. Song).

http://dx.doi.org/10.1016/j.chemolab.2016.04.002
0169-7439/

To improve the calibration performance of the random subspace ICR model, a two-level ICR model is developed in this paper. In the first level of the model, an ICA model is constructed on the original spectra data. Based on this model, various modeling directions can be determined, which are independent with each other. The diversity property of the random subspace method can be greatly improved if we build individual models along those independent directions. For construction of those subspaces, a related index is defined for variable selection in each subspace, which is based on the absolute values of the separating matrix in the ICA model. In the second level, an ICR model can be developed for each subspace. For online calibration purpose, an important issue is how to combine the results from different individual models. While one can resort to a simple average combination strategy, a more effective Bayesian based probabilistic combination strategy is proposed for results combination in this paper.

The rest of this paper is organized as follows. In Section 2, the basic ICR model is briefly introduced. Detailed demonstration of the two-level ICR model is provided in Section 3, followed by illustrations of two benchmark spectra data examples in the Section 4. Finally, conclusions are made.

2. Independent component regression (ICR)

Based on the ICA modeling method, the ICR model can be built between the extracted independent components and quality variables. In 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 $\mathbf{s} \in R^{r\times 1}$, the relationship between them is given by [22]

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

where $\mathbf{A} \in R^{m\times r}$ is the mixing matrix, $\mathbf{e} \in R^{m\times 1}$ is the residual vector. The basic problem of ICA is to estimate the original component $\mathbf{s}$ and the mixing matrix $\mathbf{A}$ from $\mathbf{x}$. Therefore, 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, given as

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

After the independent components have been estimated from the process data, the linear regression can be carried out between two datasets: the independent component dataset $\hat{\mathbf{S}} = [\hat{\mathbf{s}}_1, \hat{\mathbf{s}}_2, \cdots, \hat{\mathbf{s}}_n]^T \in R^{n\times r}$ and the quality variable dataset $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n]^T \in R^{n\times p}$. Therefore, the linear regression matrix can be calculated as

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

If we 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 determined as

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

3. Two-level ICR for multivariate calibration

Denote the whole variable 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 ICA decomposition can be carried out on $\mathbf{X}$, thus [22]

$$\mathbf{X} = \mathbf{A}\mathbf{S} + \mathbf{E} \tag{5}$$

where $\mathbf{A}$ is the mixing matrix of the ICA model, $\mathbf{S}$ is the data matrix of the independent components, $\mathbf{E}$ is the residual matrix. The number of independent components k can be determined by the negentropy method, non-Gaussianity measurement, etc. Based on the independent behavior of the extracted components, a subspace can be defined through each independent component direction, which are orthogonal to each other. Therefore, to build the ICR model in each subspace, the importance of each variable in different subspaces should be measured, depending on which the most informational ones should be retained in their corresponding subspace. To this end, an independent component related index (RI) is defined as follows

$$RI(i,j) = \frac{|w_{ij}|}{|w_{i1}| + \cdots + |w_{ij}| + \cdots + |w_{im}|} \tag{6}$$

where $i = 1, 2, \cdots, k, j = 1, 2, \cdots, m, w_{ij}$ is the j-th element of the i-th independent component direction in the separating matrix $\mathbf{W}$. Therefore, the larger the value of the j-th element, the more significant contribution it has provided through the i-th independent component direction. Based on this defined index, the importance values of different process variables through each independent component direction can be measured and ranked from the most important one to the least important one. An appropriate number of variables can be selected to form each subspace, depending on the selection scheme. The ICR model-based subspaces can be represented as follows

$$\mathbf{X}\rightarrow\begin{cases}\mathbf{X}_1 = \mathbf{X}(\mathbf{S}_1)\rightarrow subspace\ \#1\\ \mathbf{X}_2 = \mathbf{X}(\mathbf{S}_2)\rightarrow subspace\ \#2\\ \vdots\\ \mathbf{X}_k = \mathbf{X}(\mathbf{S}_k)\rightarrow subspace\ \#k\end{cases} \tag{7}$$

where $\mathbf{S}_i, i = 1, 2, \cdots, k$ is the column vector, which related to each subspace along the corresponding IC direction. A diagram of the proposed subspace modeling approach is given in Fig. 1.

3.1. ICR modeling in each subspace

Suppose the whole variable set has been divided into k subspaces, and m_b variables have been selected in each subspace, where $b = 1, 2, \cdots, k$, the corresponding dataset for each subspace can be represented as $\{\mathbf{X}_b\}_{b=1,2,\cdots,k}$. Denote the quality variable dataset as $\mathbf{Y} \in R^{n\times p}$, the subspace ICR model can be constructed as follows

$$\mathbf{X}_b = \mathbf{A}_b\mathbf{S}_b + \mathbf{E}_b \tag{8}$$

$$\mathbf{Q}_b = \left(\mathbf{S}_b^T\mathbf{S}_b\right)^{-1}\mathbf{S}_b^T\mathbf{Y}_b. \tag{9}$$

Therefore, the regression form of the subspace ICR model can be developed as

$$\mathbf{Y}_b = \mathbf{Q}_b^T\hat{\mathbf{S}}_b = \mathbf{Q}_b^T\mathbf{W}_b\mathbf{X}_b = \mathbf{R}_{ICR,b}\mathbf{X}_b. \tag{10}$$

3.2. Online calibration based on two-level ICR model

Based on the developed subspace ICR models, the property value of the new spectra data can be calculated by combing the results obtained in different subspaces. Therefore, when the new data sample $\mathbf{x}_{new} \in R^m$ is available, each subspace ICR prediction result is calculated in the first step, given as

$$\hat{\mathbf{y}}_{new}^b = \left(\mathbf{R}_{ICR,b}\right)^T\mathbf{x}_{new}(\mathbf{SI}_b) \tag{11}$$

where $b = 1, 2, \cdots, k$, $\mathbf{R}_{ICR,b}$ the regression matrix of the b-th subspace ICR model, $\mathbf{SI}_b$ represents the variable index of each subspace in the original variable space. When all of the subspace prediction results have been generated, the next step is to combine them in a certain