Analytica Chimica Acta 925 (2016) 16–22

Contents lists available at ScienceDirect

Analytica Chimica Acta

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

High and low frequency unfolded partial least squares regression based on empirical mode decomposition for quantitative analysis of fuel oil samples☆

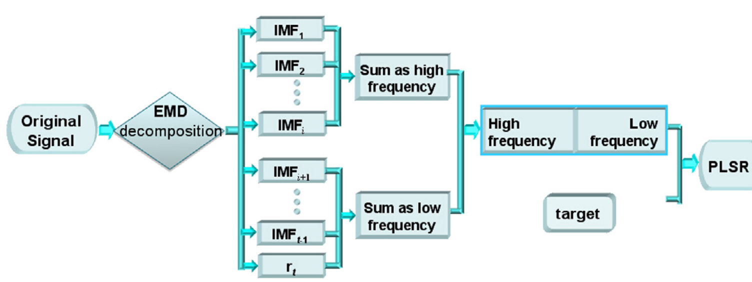
Xihui Bian [a, b, *], Shujuan Li [b], Ligang Lin [a], Xiaoyao Tan [a], Qingjie Fan [c], Ming Li [b]

[a] State Key Laboratory of Separation Membranes and Membrane Processes, Tianjin Polytechnic University, Tianjin, PR China
[b] School of Environmental and Chemical Engineering, Tianjin Polytechnic University, Tianjin, PR China
[c] Tianjin Green Security Technology Co. Ltd, Tianjin, PR China

HIGHLIGHTS

- A novel regression model integrating advantages of EMD, unfolded strategy and PLSR is proposed for the quantitative analysis of fuel oils.
- EMD and unfolded strategy are introduced for generation and integration of the member models, respectively..
- PLSR model is built between the extended dataset and the target values.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 26 August 2015
Received in revised form
31 March 2016
Accepted 21 April 2016
Available online 25 April 2016

Keywords:
Empirical mode decomposition
Unfolded strategy
Partial least squares regression
Ensemble modeling
Complex sample analysis

ABSTRACT

Accurate prediction of the model is fundamental to the successful analysis of complex samples. To utilize abundant information embedded over frequency and time domains, a novel regression model is presented for quantitative analysis of hydrocarbon contents in the fuel oil samples. The proposed method named as high and low frequency unfolded PLSR (HLUPLSR), which integrates empirical mode decomposition (EMD) and unfolded strategy with partial least squares regression (PLSR). In the proposed method, the original signals are firstly decomposed into a finite number of intrinsic mode functions (IMFs) and a residue by EMD. Secondly, the former high frequency IMFs are summed as a high frequency matrix and the latter IMFs and residue are summed as a low frequency matrix. Finally, the two matrices are unfolded to an extended matrix in variable dimension, and then the PLSR model is built between the extended matrix and the target values. Coupled with Ultraviolet (UV) spectroscopy, HLUPLSR has been applied to determine hydrocarbon contents of light gas oil and diesel fuels samples. Comparing with single PLSR and other signal processing techniques, the proposed method shows superiority in prediction ability and better model interpretation. Therefore, HLUPLSR method provides a promising tool for quantitative analysis of complex samples.

☆ Selected paper from 15th Chemometrics in Analytical Chemistry Conference, 22–26 June 2015, Changsha, China.

* Corresponding author. State Key Laboratory of Separation Membranes and Membrane Processes, School of Environmental and Chemical Engineering, Tianjin Polytechnic University, Tianjin, 300387, PR China.
E-mail address: bianxihui@163.com (X. Bian).

http://dx.doi.org/10.1016/j.aca.2016.04.029
0003-2670/© 2016 Elsevier B.V. All rights reserved.

1. Introduction

The hydrocarbon contents of fuel oil (gas oil and diesel) is a key indicator for fuel oil quality, which directly influences its combustion efficiency, engine life and automotive emissions [1,2]. The internationally traditional methods for analysis of fuel oil contents are the American Society for Testing and Materials (ASTM) methods, such as Chromatographic [3,4] and Nuclear magnetic resonance (NMR) [5] methods. Most of these methods are time-consuming, expensive and challenging in practice since many steps or professional operation skills are usually involved [6,7]. Therefore, rapid and simple analytical techniques to quantify the components of fuel oil are gaining increasing attention. Molecular spectroscopic analysis technology [6–12], especially ultraviolet (UV) spectroscopy [10–12], has shown its superiority in analysis of petroleum products, since it is cheap, fast, nondestructive and no environmental pollution. However, the high overlapped spectral bands and the existence of the noise and background of spectra make quantitative analysis with univariate calibration inapplicable. Hence, multivariate calibration methods become hot points in quantitative analysis of spectra in recent years.

Many multivariate calibration methods, such as partial least squares regression (PLSR) [13–15], artificial neural network (ANN) [16], support vector regression (LS-SVR) [1,17] etc., have been applied to analyze multi-components spectroscopic data. However, the predictive performance of these traditional calibration techniques is usually unsatisfactory because only a single model is built between the spectra and targets to predict the unknown samples, which leads to the emergence of the so-called "ensemble modeling" [18–20]. The ensemble technique can achieve better accuracy than single models and gain increasing attentions in multivariate calibration. The basic idea of ensemble modeling is combining the results of multiple individual models (or member models) to produce the final prediction. Therefore, the ensemble modeling can be broken into three questions: generation, modeling and integration of the member models [21]. In the model algorithm aspect, all the single modeling technique can be used, such as PCR, PLSR, SVR, ANN, etc. Due to its simple, rapid and good predictive performance, PLSR is used in this research. Thus, generating and integrating the member models are the keys to the success of ensemble modeling.

The existing member model generation techniques mainly include sample direction and variable direction resampling methods such as bagging [18,22], boosting [5,8,19,23–25], subspace [21] and stacked [26–28], etc. These methods can improve the predictive accuracy, stability and robustness by resampling a number of samples or variables from the whole training set many times. However, most spectra collected from spectroscopic instruments are inherently local in nature and with different localizations in both time (wavelength) and frequency [29]. Both sample direction and variable direction resampling techniques are all generated sub-models from original data of time (wavelength) domain, which cannot utilize the information in both time and frequency of the signal at the same time. Wavelet transform (WT) has shown its effectiveness for multivariate calibration due to the ability of time-frequency resolution [30]. Recently, WT is introduced for generating member models by converting the original data into the wavelet space [31,32]. Nevertheless, WT cannot process molecular spectroscopy perfectly in which nonlinear phenomenon inevitably exits and need choose appropriate settings (wavelet family, scale and number of decomposition levels) for a specified application [32]. Therefore, it is necessary to develop new sub-models generation method which not only has good localization properties both in time and frequency domains but also can make up the deficiency of WT.

Empirical mode decomposition (EMD), proposed by Huang et al. [33], is a self-adaptive signal processing technique that can be applied to process non-linear and non-stationary signal perfectly. By EMD, the complicated signal can be decomposed into a finite number of almost orthogonal intrinsic mode functions (IMFs) components and a residue component according to the inherent characteristics of the signal. Because of its efficiency, the EMD method has been successfully applied in several fields such as price forecasting, biomedical engineering, earthquake engineering, tourist arrivals, electrical power system and mechanical fault diagnosis [34–36]. Up to now, few research has been reported on ensemble modeling by EMD in spectra analysis. Thus, EMD was introduced to generate member models for fully using the information embedded over frequency and time domains.

Integration of the member models is another key for ensemble modeling, which is generally implemented by combination of the results obtained by all the member models [37]. Simple average and weighted average are two commonly used integration strategies. The former does not need any parameters, but the results of which are not as good as the latter while the latter need to determine the weights for the member models. Although several criteria have been used such as prediction error [24], prediction residual error sum of squares (PRESS) [30], nonnegative least square [26] etc., it is difficult to determine the criterion which can give the optimal predictive result. Hence, the determination of weights is still a problem for a given application. Recently, Shao et al. [37] proposed wavelet unfolded partial least squares (WUPLS) for quantitative analysis of complex samples. In this method, different wavelets were used to process the measured spectra, and then all the processed data were extended in the variable direction instead of building multi-models. This unfolded strategy can circumvent the determination of model weights.

In this paper, a novel regression model that integrates EMD and unfolded strategy with PLSR, referred to the high and low frequency unfolded PLSR (HLFUPLSR), is proposed for quantitative analysis of the fuel oil samples. In order to demonstrate the enhancement of predictive accuracy by the proposed approach, PLSR with original spectra and preprocessed [38] spectra by the first order derivative (1st derivative), continuous wavelet transform (CWT) are used for comparisons. Moreover, light gas oil and diesel fuels samples were used to evaluate the performance of the method.

2. Theory and algorithm

2.1. Empirical mode decomposition (EMD)

The most appealing nature of EMD is its adaptive data-driven decomposition mechanism which does not require a priori-defined basis such as Fourier and Wavelet transform [36]. By using EMD, any complicated signal can be decomposed into a finite number of IMFs and a residue according to its inherent characteristics. Thus, owing to the self-adaptability of EMD, the IMFs number is determined by the signal itself. Fig. 1 shows the extracting process of IMF for a simulated signal by EMD. To obtain the IMFs, it can be operated by the following steps. Firstly, all the local maxima and minima are identified for the given signal $\mathbf{x}$ (black line) and then the upper (red line) and lower envelopes (green line) are formed by connect them with a cubic spline line, respectively. The mean value curve (blue line) of the two envelopes is calculated by averaging the two envelopes. The difference component $\mathbf{c}$ is gotten by subtraction the mean value from the original signal $\mathbf{x}$. An IMF IMF_i will be obtained if $\mathbf{c}$ satisfies the two constraints of IMF (a) the number of extrema in the whole data must either equal with the zero-crossings or differ at most by one; (b) the mean value of the two envelopes defined by the maxima and the minima are symmetric with respect to zero mean [34]. Then $\mathbf{x}$ is replaced with the residual $\mathbf{r} = \mathbf{x} - \mathbf{c}$. If $\mathbf{c}$ is not an IMF, $\mathbf{x}$ is replaced with $\mathbf{c}$. This process is called a sifting process and it is repeated until the residual satisfies the stopping criterion. At the end of this process, the signal can be