Application of various Pre-processing techniques on Infrared (IR) Spectroscopy data for rapid classification of different ghee samples

Navjot Kumar [1],[2], PC Panchariya[1],[2], AH Kiranmayee[1],[2], Surendra Singh Patel[2], Rishi Ranjan[1][2]
[1] CSIR-Central Electronics Engineering Research Institute, Pilani, Jhunjhunu-333031.
[2] Academy of Scientific & Innovative Research (AcSIR), Chennai, Tamil Nadu.
Email: navjot@ceeri.res.in, patelsurendra@ceeri.res.in,

Abstract—**When dealing with IR spectroscopy, the pre-processing of spectral data is considered to be one of the most important part of chemmometrics modeling. Due to any uncontrollable physical variations may lead to an additive, multiplicative and wavelength-dependent scattering effects in the recorded spectra. Pre-processing techniques basically are required to remove these scattering effects from the spectra and subsequently improve the further quantitative and qualitative analysis. Most popular pre-processing techniques are; baseline correction, smoothing of the spectra, normalization, scattering correction and spectral derivatives. This paper begins with the theoretical and mathematical foundation of various pre-processing techniques used for IR spectroscopy. Then a qualitative analysis is performed by applying these techniques on the spectral data collected using various samples of ghee. The comparison of various pre-processing is obtained by modeling of the data using Principle Component Analysis (PCA) and then the k-means clustering algorithm.**

Keywords—chemmometrics; normalization; scattering correction; spectral derivatives; PCA.

I. Introduction

When IR reflectance or transmittance mode measurements are performed on a sample, often significant differences are observed in the spectra. There can be so many parameters which can cause these scattering effects e.g uncontrollable physical variations (e.g change in the temperature), change in the refractive index, surface roughness or shape etc. These factors contribute in additive, multiplicative and wavelength-dependent scattering effects for example shifts in the baseline, tilt in the spectra, and unwanted peaks at some instances for certain wavelength region in the spectra. Sometimes these scattering effects put such an impact over the spectra that spectral measurement for a given chemical composition can sometimes look completely different [1][3][4]. Problem becomes even much more complicated when the system has to distinguish between the spectra having very small variations. As mentioned above, the scattering effect in NIRS mainly consists of; Additive effect (shift in the baseline), Multiplicative effect (scaling of the entire spectra), frequency dependent baseline variations [1][2][4]. All these effects can be linear or non-linear, which makes it very difficult to minimize or remove from the spectra. These unwanted variations are independent of any chemical response and may lead to inaccurate results, while performing qualitative or quantitative chemical analysis.

Therefore it is very important to apply appropriate pretreatment techniques to the recorded spectra before going for any data analysis. The success of the pretreatment lies when it is able to perform mathematical treatment by which one can separate the light scattering from light absorbance, transmittance or reflectance in the spectra. The most commonly used pre-processing data treatment methods are multiplicative scatter correction (MSC), extended multiplicative signal correction (EMSC), standard normal variate (SNV), de-trending, spectral derivatives, orthogonal signal correction (OSC) and generalized least square [1] etc. The data scientists should apply out single or combinations of these different pre-treatment methods to the spectra and assess which of the methods are giving better results by comparing subsequent model performance. However, the prior knowledge of these pre-treatment methods and the collected spectra is very important as the random trials of these algorithms may cause damage to the signal of our interest.

In the following sections, the various popular pre-treatment methods will be described and evaluated by applying these to the spectral data for different ghee samples collected from MIR spectroscopy technique. The evaluation is basically done by performing the pre-treatment first followed by modeling of the data using Principle Component Analysis (PCA) and then the k-means clustering algorithm. Collectively, this work can help in better understanding of various data pre-treatment methods and their effects on the recorded spectra from MIR spectroscopy.

II. Materials and Methods

A. Collection of various ghee samples

For this work total eight ghee samples were collected, two were made at home itself using cow's & buffalo's milk, rest of the five were directly purchased from the market as packaged ghee. Out of the packaged ghee two brands are from cow's milk, two from buffalo's milk and one sample is vanaspati (dalda) ghee. we have given different codes to these brands as shown in the Table.1. Before performing measurements, we have taken samples from each of these brands and separately melted these by heating up to 50^0C. The samples were kept at this temperature for around 30 minutes and then the temperature was reduced to 40^0C. Every time 2ml of the sample was taken and measurement was performed using MIR spectrophotometer. The experimental setup design is shown in the Fig. 1.