ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

# A Study on Various Preprocessing Algorithms Used For NIR Spectra.

**A Anne Frank Joe[1], and A Gopal[2].**

[1]Research Scholar, Sathyabama University (Electronics and Instrumentation Department), Jeppiaar Nagar, Old Mamallapuram Road, Chennai 119, India.
[2]Sr. Principal Scientist, CEERI – CSIR, Taramani, Chennai 113, India.

**ABSTRACT**

NIR spectroscopy is one of the most common non-destructive testing techniques with fingerprint spectra of each compound for swift characterization of samples. Chemometrics is the discipline utilizing mathematical and statistical process to draw a clear picture of the quantitative results of the chemical constituents like protein, carbohydrates, fat, moisture, sugar content etc and to relate its quality properties to analytical instrument data. A basic model for the chosen data is obtained and these blueprint models can be consistently applied to future data in order to calculate similar quality parameters. The fundamental and first step in chemometrics modeling is pre-processing of the acquired spectral data. The goal of pre-processing is to eradicate possible noise or disturbance in the spectra. Thus enabling the variation due to the chemical constituents to be the sole focus that is to be elaborated to study the sample in detail, thereby the subsequent multivariate regression is fine tuned into refined classification model. There are many different types of data pre-processing beginning from standard normal variate transformation (SNV), Savitzky-Golay first derivative transformation and wavelet transforms (WT) and other simple smoothening functions on the NIR spectra. The selection of a suitable pre-processing technique depends on the nature of the signal. While obtaining NIR spectra of solid samples, variation may generally be caused by chemical composition of the sample, scatter of radiation at the surface of particles and the irregular particle size causing spectral path length to differ through the sample. The diverse pre-processing techniques practiced and their qualitative and quantitative impact on their end applications are analyzed to provide NIR users with better end models. The aim is to provide NIR users with refined and better performing models through fundamental knowledge on spectral pre-processing.

**Keywords:** Pre-processing, NIR, Chemometrics, SNV, Savitzky-Golay

***Corresponding author***

ISSN: 0975-8585

## INTRODUCTION

The method of non destructive measurement of the internal attributes of a compound is most welcomed in various industry, where quality evaluation is an important issue. The usual procedures available to measure these qualities are based on complex and time consuming processes which involves expensive chemical reagents, also consuming substantial amount of manual labor and yet destructive. It is always profitable to develop non- destructive methods for measuring the internal attributes which performs much efficient in lesser time. Near infrared spectroscopy(NIRS) is one such analytical technique. One of the many advantages of NIRS is that it enables to study several constituents simultaneously. However, NIRS provides spectral data that is polluted with noise and is influenced by a number of physical, chemical, and structural variables

### Sources of error

NIR technique can be applied to samples of different states. The difference in the size of the particle will cause scattering and contributes to variation in NIR spectra. Such variations might have an additive and multiplicative effect on the sample spectra. Additive effects cause vertical shift whereas multiplicative effects result in difference in slope for spectra acquired when compared to a reference spectrum. The non uniform particle size of solids causes the NIR diffuse reflectance spectrum to be corrupted by scattering noise, scattering also occurs on the top layer of a material. Samples belonging to the same batch may have spectral differences among them due to the non-uniform particle size. The size dictates the path length of reflectance or transmittance, variations in the size may contribute to baseline shift in the spectra.

### Correction methods

Amid the several pre-processing techniques in NIR spectroscopy, the most commonly practiced methods can be broadly classified into two categories namely, the method for scatter correction and spectral derivatives.

*The scatter-correction* methods comprises of Multiplicative Scatter Correction (MSC), Extended MSC (EMSC), Inverse MSC, Extended Inverse MSC, Standard Normal Variate (SNV) , de-trending, and normalization[1].

*The spectral derivative* methods namely, (SG) Savitzky-Golay polynomial derivative and Norris-Williams (NW) derivative filters are also practiced. Smoothing of the spectra before the calculation of the derivative is generally advised, this decreases the ill effects on the signal-to-noise ratio.

### Real data set

The spectrum on which we apply our procedure is built from 165 near infra red reflectance spectra of whole wheat grain samples measured on a FOSS XDS OptiProbe Analyzers with wavelength range of 400nm-2500nm. One set of five raw spectra obtained from a single sample set are depicted in Figure-1a .The best preprocessing method varies from sample to sample, currently the suitable method applied is based on trial and error. The design of a methodology with a efficient and organized approach would be a boon. In reality it is possible to have a few route to indication to decide the choice of pre-processing. Example, the figure shows five spectra of the same wheat sample UP 262. The spectrum was obtained by placing the probe at different positions on the same sample. The Chemical components being the same in the sample, all the five spectrum should have traces only one spectra due to overlap of the spectrum. It is observed that the spectrum are spread apart and not overlapping. An ideal pre-processing approach will result in minimum deviation between replicates. This may not be the most accurate indicator for pre-processing selection, this is one of the good practices.

### Data Pre-treatment Methods in detail

### Multiplicative Scatter Correction (MSC)

When dealing with powder or solid sample spectra, physical light-scattering effect pollute the light absorbance due to chemical components. With respect to overcome the intervention of light scattering,

multiplicative scatter correction method is an ideal and efficient choice in the pretreatment process of spectral data. This techniques can isolate the actual data with details of chemical absorbance of the sample wheat grain from the interfering scattering signal in the data[2].

The light scattering or change in path length for each sample is estimated relative to that of an ideal sample. A close fit is created between the ideal and the mean spectrum by using least squares.

$$X_{i=}A_{i+}B_iX_j+E_i$$

$X_i$ is the individual spectrum I, $X_j$ is the mean spectrum of the spectrum
$E_i$ is the actual chemical information in the spectrum i.
$A_i$ is the intercept and $B_i$ is the slope.

$$X_{msc,i}= (X_i - A_i)/ B_i$$

Where $X_{msc,i}$ is the MSC corrected spectrum

The ideal sample is generated from the mean of the original spectrum. MSC techniques is performed by cross checking the original raw spectrum against a reference spectrum, the raw spectrum is then corrected by utilizing the slope and intercept of the linear fit . MSC has been proved to be an effective method in minimizing the multiplicative effect and baseline shifts. The result of MSC is quite similar to that of SNV[6], it is clearly evident at a glance at figure outputs of MSC and SNV here. SNV corrects each spectrum individually unlike MSC which needs the entire data set. The flow of process begins with estimating the spectral portion which is influenced solely by the light scattering and does not contain any chemical information.In practice identifying such an area is difficult, also such a spectral portion will have a poor SNR value. Thus the entire spectrum is considered most of the times. MSC performance can be enhanced further by performing an offset correction in advance.

The baseline shift of the spectrum is corrected, The spreading of spectrum is also reduced and the 5 spectra are made to overlap to a considerable extent. MSC is dependent on the raw spectrum collection, if there is any change in the original spectrum, the MSC has to be recalculated.

**Extended Multiplicative Signal Correction (EMSC)**

The multiplicative scatter correction (MSC) was further developed into extended multiplicative scatter correction (EMSC) in the year 1989 by Stark and Martens. . Extended Multiplicative Scatter Correction (EMSC) is a powerful preprocessing technique to separate and eliminate complex multiplicative and additive effects. This technique is specifically useful in minimizing light scattering variation that are wavelength-dependent[5],[6] . The EMSC method employs the pure spectra of the wheat grain sample and interference effects to improve the optical path length estimation.

From the raw spectral data, a mean spectra is calculated. A least square fit is devised between mean spectrum and each of the raw spectrum thus creating a first-degree polynomial. One spectrum at a time is corrected by using the mean spectrum. A least square fit is again devised between the mean spectrum and each of the new spectrum (first-degree polynomial). This corrects the new spectrum.

After pre-treatment the corrected spectra becomes insensitive to light scattering variations and responds linearly to the analyte concentration. However, EMSC cannot be wildly used due to a lack of the pure spectrum of chemical substance under consideration. In general EMSC has a small effect on baseline and the Detrend is comparatively better.

**Standard Normal Variate (SNV)**

SNV is a very simple though effective pre-treatment method for scattering correction. The five spectra which should have ideally overlapped have non identical spectra with added baseline shift as well. SNV is a suitable choice for this spectrum needing correction. Technically this is similar to auto scaling the rows instead of the columns in the matrix which builds the spectrum.

ISSN: 0975-8585

The multiplicative variations between spectra which needs to be corrected are caused from accidental differences in path length of the sample. The path length variation originates due to differences in sample particle size, thickness, preparation and presentation[2]. Correction of multiplicative effect is not as easy as applying derivatives or scaling, centering or scaling. The transformation is performed on each spectrum thereby it will obtain an absorbance spectrum with mean 0 and standard deviation 1.The SNV corrected spectrum is calculated as

$$Xi,m^{SNV}=(X_{i,m}-X_i^{'})/S_i$$

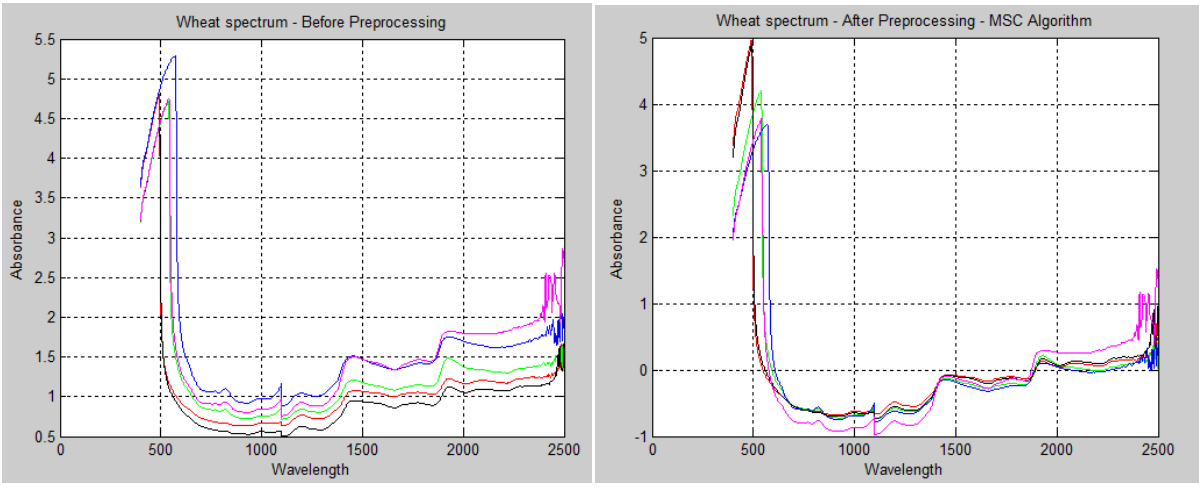
i.e, The SNV transformation is applied to each raw spectrum by subtracting the spectrum mean and scaling with the spectrum standard deviation.
$Xi,m^{SNV}$ is the absorbance value after the pretreatment.
Xi,m is the absorbance value of raw spectrum and
$S_i$ is the standard deviation.
Standard normal variate (SNV) transformation removes the slope variation from spectra[4].

**Derivation**

*Savitzky-Golay (SG) smoothing* is a commonly used method of preprocessing that can effectively eliminate the noises like baseline-drift, tilt, reverse[1]. It contains many different smoothing modes. The smoothing parameters include the polynomials degree (PD): the larger the window and lower the polynomial order, the more smoothing that occurs, the derivatives order of polynomials (DOP), and number of smoothing points (NSP): very little NSP could cause calculation error, resulting in a decreased model precision, while a large NSP would over smooth and buff the spectral data, leading to the decreased accuracy. A careful choice on the number of smoothing points is very important in SG smoothing. A set of consecutive pints on the spectra is chosen and a least square fit to a polynomial is done. The central point of the newly fitted polynomial curve is considered as the new smoothed point on the spectrum.

*Savitzky-Golay (SG) Derivative,* a set of integers (x1, X2,…$x_n$) can be derived from the spectral data and they can be employed as weighting coefficients for the smoothing function. This is very similar to fitting the data to a polynomial, only that this method is much more effective in computation and execution is much faster. Derivatives mimic high-pass filter and is frequency-dependent. The method is suggested to be employed where the variables are strongly related to each other and consecutive variables contain quite similar correlated data[6].

The point-difference method of first derivative is quite simple, here each point in a spectrum is subtracted from its adjacent spectral point. This retains the signal portion that is different and nullifies the similar signal points in the spectra. The same is repeated on the complete spectra thereby removing any offset from the sample and reduces the lower-frequency signals.



ISSN: 0975-8585

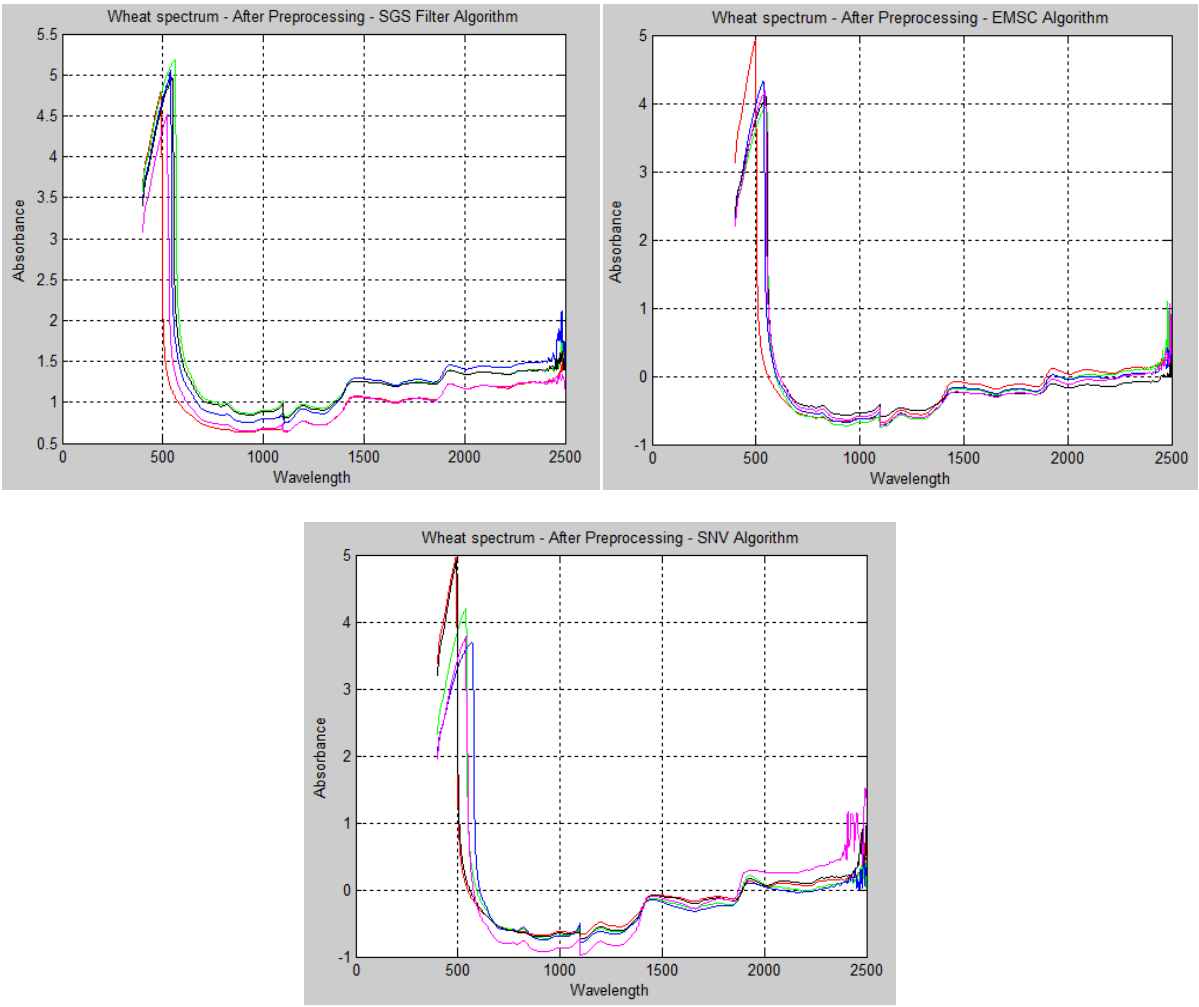



**Figure (a), (b),(c),(d),(e)- Plot of wavelength and absorbance of the 5 spectra taken from the same sample -Before preprocessing, On applying MSC algorithm, SGS algorithm, EMSC algorithm and SNV algorithm.**

## RESULTS AND DISCUSSION

Thirty wheat grain samples were taken. Five spectra were taken from each sample set. Each set of five spectra were analyzed individually by applying the various preprocessing methods. From the overall data set, MSC and SNV appeared to behave similar. EMSC performed better than MSC. A total of thirteen wheat grain samples were analyzed by conventional method to measure the moisture content in the wheat grain. 250 gms of each sample was used and the test method used was IS:1155-968(RA-1994).

Five spectra from each sample had a spreading effect and baseline shift as well. SG, SNV, MSC, EMSC methods were applied on each spectra and the pretreated spectra was used to quantify the moisture content. The pretreatedspectra produced a result very closely matching the wet analysis method, whereas moisture quantification using the raw spectra showed considerable error. Thereby creating a question on the performance of the end classifier to be designed. Thus Pretreatment is an essential first step in chemometric mathematical modeling. A combination of different pre treatment methods can also be applied to obtain the best results of removing the interfering information.

## REFERENCES

[1] http://www.americanpharmaceuticalreview.com/Featured-Articles/116330-Practical-Considerations-in-Data-Pre-treatment-for-NIR-and-Raman-Spectroscopy/
[2] Verboven, S., Hubert, M. and Goos, P. J Chemometrics 2012;26: 282–289.
[3] Moghimi, Ali, et al. Biosystems engineering 2010;106(3):295-302.

[4] Randolph, Timothy W. Cancer Biomarkers 2006;2(3):135-144.
[5] Li, Qingbo, Qishuo Gao, and Guangjun Zhang. Journal of Spectroscopy 2013.
[6] Rinnan, Åsmund, Frans van den Berg, and Søren Balling Engelsen. TrAC Trends in Analytical Chemistry 2009;28(10):1201-1222.
[7] Guidetti, Riccardo, Roberto Beghi, and Valentina Giovenzana. Chemometrics In Food Technology. INTECH Open Access Publisher, 2012.
[8] Alander, Jarmo T., et al. International Journal of Spectroscopy 2013.