Insights into information contained in multiplicative scatter correction parameters and the potential for estimating particle size from these parameters

Yi-Chieh Chen and Suresh N. Thennadil*

Department of Chemical and Process Engineering,
University of Strathclyde, Glasgow, United Kingdom

*Corresponding Author:

Address:

75 Montrose Street,
James Weir Building,
Chemical and Process Engineering,
University of Strathclyde,
Glasgow, G1 1XJ,
United Kingdom.

Tel: +44 141 548 2241

Fax: +44 141 548 2539

Email: suresh.thennadil@strath.ac.uk

ABSTRACT

Empirical preprocessing methods such as multiplicative scatter correction (MSC) and extended multiplicative scatter correction (EMSC) are widely used to remove light scattering effects from spectra of samples containing particulate species. When these methods are used, the parameters that are applied for correcting the spectra are normally discarded. If the scatter correction method is effective, these parameters should contain information regarding the particulate species since it is this component which contributes to the light scattering effects. This study had two objectives. The first objective was to examine the nature and extent of information contained in scatter correction parameters. The second objective is to examine whether this information can be effectively extracted by proposing a method to obtain particularly, the mean particle diameter from the scatter correction parameters. The approach used for this investigation is to examine the scatter correction parameters in terms of the information regarding particle size and particle concentration by using a dataset in which particle size and particle concentration vary significantly. It was found that the MSC parameters contained significant information regarding particle size and concentration. A two-step method to obtain simultaneously the particle concentration and particle diameter was proposed and tested using a 2-component and 4-component data set. It was found that the approach which uses the MSC parameters gave a better estimate of the particle diameter compared to using Partial Least Squares (PLS) regression for the 2-component data. For the 4 component data it was found that PLS regression gave better results but further examination indicated this was due to chance correlations of the particle diameter with the two of the absorbing species in the mixture.

1. Introduction

Multivariate calibration methods such as Partial Least Squares (PLS) regression have been widely used to build calibration models for predicting the concentrations of chemical components from near-infrared (NIR) spectra. When samples containing particles are encountered, multiple light scattering effects introduce nonlinearities leading to degradation in model performance. Several empirical preprocessing methods such as multiplicative scatter correction (MSC), standard normal variate (SNV), extended multiplicative scatter correction (EMSC), orthogonal signal correction (OSC), and optical path length estimation and correction (OPLEC) have been used to mitigate light scattering effects.[1-6] When dealing with particulate systems, it is generally assumed that the information removed from the measured spectra by the application of these empirical methods is essentially the manifestation of the underlying physics of light scattering without significant loss of chemical information, thus improving the performance of the multivariate regression models in estimating chemical information from the corrected spectra.

When these methods are used, the parameters that are applied for correcting the spectra are normally discarded since they are supposed to contain only physical information. If the scatter correction method is effective, the scatter correction parameters would be expected to contain information regarding the particulate species since it is this component which contributes to the light scattering effects. If this information can be extracted then it could provide valuable extra information (particle size) in addition to estimates of concentrations which are obtained from the calibration models built on the scatter-corrected spectra.

Several studies can be found in the literature where scatter correction techniques are applied and compared in terms of the improvement in performance of models built using the corrected spectra. However, the performances of the empirical methods appear to be dependent on the system studied with no single empirical scatter correction method consistently outperforming others across a number of different types of datasets. Among the empirical methods, the more recently developed OPLEC method

68 has been promising,[6, 7] though it has not yet been applied widely enough to conclude that the method
69 is indeed consistently superior to other available methods. A study based on simulations using a
70 rigorous light propagation model indicated that most of the common scatter correction methods led to
71 similar model performances.[8] In addition, this study also indicated that the effectiveness of a
72 particular scatter correction technique was also dependent on measurement configuration. To-date
73 however, to our knowledge, there have been no in-depth studies that have examined the information
74 contained in the scatter correction parameters themselves. Such a study will be useful for understanding
75 the nature and characteristics of information contained in the parameters of a particular scatter
76 correction method. This could help in identifying situations where they perform the best and could
77 potentially help in modifying the methods to produce more effective scatter correction techniques.

78 The implicit assumption when applying scatter correction methods is that light scattering effects
79 manifesting as an additive or multiplicative or more complex (e.g. wavelength dependent) effects in the
80 measured spectra are removed. However, there are other non-chemical effects which can lead to similar
81 manifestations in the spectra as the assumed effect of light scattering (e.g. instrument drift). In other
82 words, the corrections are not necessarily specific to scattering. Hence the terms Multiplicative Signal
83 Correction and Extended Multiplicative Signal correction can sometimes be found in the literature
84 where “signal” is used instead of “scatter” to denote that the techniques are more general in terms of the
85 non-chemical information removed by them.[5] Similarly, the SNV method is clearly a general method
86 which has also been used to correct light scattering effects.

87 In any dataset consisting of spectroscopic measurements of particulate systems, we can expect the
88 non-chemical variations to be a combination of effects with the light scattering effects usually being the
89 most dominant. There are four possibilities why one scatter correction technique might work better than
90 others: (1) The method removes the most amount of variation due to light scattering compared to others;
91 (2) The method removes the most amount of variation due to all non-chemical effects present in the
92 measurements; (3) The method linearizes the measurements most effectively compared to other

93 methods, (4) The method removes the least amount of relevant chemical information; and (5) The
94 method is the most effective in terms of a combination of the previous four aspects. Therefore the most
95 effective “scatter correction” method will differ from one system to another depending on the dominant
96 type of non-chemical variations in the measurements that form the datasets.

97 This study had two objectives. The first objective was to examine the nature and extent of information
98 contained in scatter correction parameters. The second objective is to examine whether this information
99 can be effectively extracted by proposing a method to obtain particularly the particle size from the
100 scatter correction parameters. The approach used for this investigation is to examine the scatter
101 correction parameters in terms of the information regarding particle size and particle concentration by
102 using a dataset in which particle size and particle concentration vary significantly and where the values
103 of these parameters have been accurately measured. Since particle concentration and size are the two
104 sample parameters that affect the extent of light scattering by a sample, it follows that any effective
105 correction step will contain information regarding these two sample parameters. Following this logic, if
106 the scatter correction step is effective, then it should be possible to extract information regarding particle
107 size and/or particle concentrations from the scatter correction parameters. This is investigated through
108 an approach for building models to obtain particle size information using the scatter correction
109 parameters. The investigation into the effectiveness of the scatter correction approach to specifically
110 provide information regarding particle size was carried out using two models systems namely, a two
111 component and a four component system both containing polystyrene latex particles as the scattering
112 species.

114 **2. Materials and Methods**

115 **2.1 Experimental dataset**

116 The two datasets used in this study were obtained from previously published works.[9, 10] A brief
117 description of the datasets is given here. Both datasets contain measurements taken using a Cary 5000

118 spectrometer equipped with an external diffuse reflectance accessory and 1 mm sample thickness was
119 chosen. The first dataset is a polystyrene-water system that consists of a total of 35 samples with 5
120 particle diameters ($d_p = 100, 200, 300, 430$ and 500 nm) and 7 particle concentrations ($y = 0.1, 0.5, 0.9,$
121 $1.23, 1.6, 1.95$ and 2.3 in wt. %) for each particle size.[9] Spectra were collected using 0.4 sec as
122 integrating time for a wavelength range of $\lambda = 1550 - 1850$ nm with 4 nm interval, resulting in 75
123 discrete wavelengths per spectrum. The raw spectra were smoothed using Savitsky-Golay filter with
124 window width of 9 and polynomial order of 3 to remove noise in the measurements.

125 The second dataset is a 4-components system that consists of water (H_2O), deuterium oxide (D_2O),
126 ethanol (C_2H_5OH), and polystyrene particles.[10] The concentration of each component was varied so
127 that the correlation between concentration of polystyrene particles and other components in the sample
128 is negligible. In this dataset there are samples containing the same particle diameter and particle
129 concentration while concentrations for other components vary. 5 particle diameters ($d_p = 100, 200, 300,$
130 430 and 500 nm) and 5 concentrations ($y = 1, 2, 3, 4,$ and 5 in wt. %) were employed to form this dataset
131 of 45 samples. Spectra were collected in the range of $\lambda = 1500 - 1880$ nm with 2 nm intervals and 10 sec
132 as the integrating time. The same smoothing conditions applied to the first dataset were also employed
133 for this dataset before subjecting to scatter correction methods. Both datasets contained measurements
134 from three different measurement configurations namely, total reflectance (R_d), total transmittance (T_d)
135 and collimated transmittance (T_c).

136 2.3 Estimation of particle size from MSC parameters

137 The first step in this approach is to establish the relationship between the MSC parameters and
138 particle size (diameter) using the calibration dataset. In other words we develop models for expressing
139 the additive (a) and multiplicative (b) term of MSC parameters as a function of particle diameter (d_p)
140 and particle concentration (y). As will be seen in the next section, the MSC parameters are dependent
141 on both particle diameter and concentration. Given these “direct” relations, we can then write inverse
142 relations i.e. particle diameter as a function of particle concentration and MSC parameters a or b or

both. This relationship can then be used to estimate the diameter of particles in a sample i given the concentration of particles and the MSC parameters a_i and b_i for that sample. Usually the actual particle concentration of a sample is also unknown. Therefore it has to be estimated. This can be done in the usual manner of building a calibration model for the concentration using PLS regression. Then in the inverse expression, the estimated particle concentration ($\hat{y}$) is used. The methodology is summarized by the flowchart shown in Fig. 1.

The methodology consists of two stages, the calibration model building stage (Stage 1 shown in black) where the models for estimating d_p and y are developed using the calibration dataset, and the prediction stage (Stage 2 shown in blue) to estimate particle diameter $\hat{d}_p$ and particle concentration $\hat{y}$ from spectra of unknown sample conditions. for a two component system is considered. In Stage 1, measured spectra ($\mathbf{x}_{meas}$) from a set of calibration samples of known y and d_p is subjected to an empirical scatter correction method such as MSC. The MSC equation is given by:

$$\mathbf{x}_{meas} = a + b\mathbf{x}_{ref} + \mathbf{e} \quad (1)$$

where $\mathbf{x}_{meas}$ is the spectrum measured from the sample, and $\mathbf{x}_{ref}$ is a reference spectrum. The values of parameters a and b are estimated using ordinary least-squares regression of $\mathbf{x}_{meas}$ onto $\mathbf{x}_{ref}$. The error term, $\mathbf{e}$, contains the chemical information of the sample since it is the portion that is not explained by the physical variations (changes in baseline/slope). Note that the letters in bold indicate vectors. Once a and b are estimated, Eq. (1) can be rearranged as follows:

$$\mathbf{x}_{corr} = (\mathbf{x}_{meas} - a)/b = \mathbf{x}_{ref} + \mathbf{e}/b \quad (2)$$

where $\mathbf{x}_{corr}$ is the spectrum corrected using MSC and should be as similar to $\mathbf{x}_{ref}$ as possible (in a least squares sense). This means that the difference between $\mathbf{x}_{corr}$ and $\mathbf{x}_{ref}$, i.e. $\mathbf{e}/b$, can be considered to be independent of the scattering effect. In this work, the reference spectrum for this example was taken to be the average spectrum of the whole calibration dataset.

Based on the functional forms identified through the analysis of the relationships between the MSC parameters and the particle diameter (d_p) and concentration (y) are obtained. For the two-component dataset, the expressions were (discussed in §3.1):

$$a = \xi_a + (\alpha_1 y + \alpha_2 y^2 + \alpha_3 y^3) (1 + \beta_1 d_p + \beta_2 d_p^2) \quad (3)$$

$$b = \xi_b + \kappa_1 y (1 + \eta_1 d_p + \eta_2 d_p^2) \quad (4)$$

where coefficients (α , β and η) were determined based on the best fit of y and d_p to the MSC parameters a and b . It is worth noting that the expressions may not be unique therefore care has to be taken to ensure that the coefficients used in the functional forms are significant.

Eqs. (3) and (4) can then be re-arranged so that d_p can be expressed as:

$$d_p = f(a, y) = \frac{1}{2\beta_2} \left\{ -\beta_1 \pm \left[\beta_1^2 - 4\beta_2 \left(1 - \frac{a - \xi_a}{\alpha_1 y + \alpha_2 y^2 + \alpha_3 y^3} \right) \right]^{1/2} \right\} \quad (5)$$

$$d_p = g(b, y) = \frac{1}{2\eta_2} \left\{ -\eta_1 \pm \left[\eta_1^2 - 4\eta_2 \left(1 - \frac{b - \xi_b}{\kappa_1 y} \right) \right]^{1/2} \right\} \quad (6)$$

It is also possible to obtain an expression for d_p that includes both parameters a , b and the measured and corrected spectrum, $\mathbf{x}_{meas}$ and $\mathbf{x}_{corr}$. The expression simultaneously makes use of particle size information contained in these parameters as well as that remaining in the corrected spectrum, thereby providing the possibility of better estimation of d_p owing to the augmented information contained in such an expression. In order to do this, we start with the re-arranging Eq.(2):

$$\mathbf{x}_{meas} = a + b\mathbf{x}_{corr} \quad (7)$$

Substituting Eqs. (3) and (4) into Eq. (7), and carrying out algebraic manipulations an expression for d_p as a function of a , b , $\mathbf{x}_{corr}$, $\mathbf{x}_{meas}$ and y can be obtained. Maple version 13 (Waterloo Maple Inc.) was employed to solve for d_p to obtain the following expression:

$$d_p = h(a, b, y, \mathbf{x}_{meas}, \mathbf{x}_{corr})$$

$$186 \quad = \frac{-1}{2y(\alpha_3\beta_2y^2 + \alpha_2\beta_2y + \alpha_1\beta_2 + \kappa_1\eta_2\mathbf{x}_{corr})} \left\{ \pm \left[\frac{\left(\alpha_3\beta_1y^3 + \alpha_2\beta_1y^2 + \alpha_1\beta_1y + \kappa_1\eta_1y\mathbf{x}_{corr} \right)^2}{\left(\alpha_3\beta_2y^2 + \alpha_2\beta_2y + \alpha_1\beta_2 + \kappa_1\eta_2\mathbf{x}_{corr} \right)} \right]^{1/2} \right\} \quad (8)$$

187 Note that $\mathbf{x}_{corr}$ and $\mathbf{x}_{meas}$ are scalars when writing d_p in this form indicating that the measured and
 188 corrected absorbance in the equation are for a particular wavelength. Therefore we obtain a solution for
 189 d_p at each wavelength. As a result d_p estimated by this equation is obtained by averaging over all the
 190 wavelengths.

191 For the 4-component data, following the same procedure leads to the following equations.

$$192 \quad a = \xi_a + \alpha_1y + \alpha_2y^2 + \beta_1d_p + \beta_2d_p^2 + \gamma_1y \cdot d_p + \gamma_2y \cdot d_p^2 + \gamma_3y^2 \cdot d_p + \gamma_4y^2 \cdot d_p^2 \quad (9)$$

$$193 \quad b = \xi_b + \kappa_1y + \kappa_2y^2 + \eta_1d_p + \eta_2d_p^2 + \varsigma_1y \cdot d_p + \varsigma_2y \cdot d_p^2 + \varsigma_3y^2 \cdot d_p + \varsigma_4y^2 \cdot d_p^2 \quad (10)$$

$$194 \quad d_p = \frac{-b \pm \sqrt{b^2 - 4a \cdot c}}{2a} \quad (11)$$

$$195 \quad \text{where } a = \beta_2 + \gamma_2 \cdot y + \gamma_4 \cdot y^2 + x_{corr} \cdot (\eta_2 + \varsigma_2 \cdot y + \varsigma_4 \cdot y^2)$$

$$196 \quad b = \beta_1 + \gamma_1 \cdot y + \gamma_3 \cdot y^2 + x_{corr} \cdot (\eta_1 + \varsigma_1 \cdot y + \varsigma_3 \cdot y^2)$$

$$197 \quad c = \xi_a + \alpha_1 \cdot y + \alpha_2 \cdot y^2 - x_{meas} + x_{corr} \cdot (\xi_b + \kappa_1 \cdot y + \kappa_2 \cdot y^2)$$

198
199

200 In Stage 2, the spectrum of a sample whose particle size and concentration have to be estimated is
 201 subjected to the scatter correction method using the same reference spectrum ($\mathbf{x}_{ref}$) that corrects the
 202 calibration set. The corrected spectrum is then subjected to the PLS calibration model built in Stage 1 to
 203 obtain an estimate of the particle concentration $\hat{y}$. This value of $\hat{y}$ is then used along with one of the

three inverse expressions mentioned above to get $\hat{d}_p$. Thus estimates for both particle diameter and concentration are obtained from the spectrum.

It should be noted that while the methodology is described for the case where MSC is used as scatter correction method, it can be easily applied to any other scatter correction technique provided the scatter correction parameters obtained from a technique have extractable information regarding the particle diameter.

3. Results and Discussion

An initial analysis was carried out using data from each of the measurement configurations, namely total transmittance (Td), total reflectance (Rd) and collimated transmittance (Tc). MSC, and two versions of EMSC namely EMSCL and EMSCW [8, 11] were applied to the datasets and the scatter correction parameters were examined. In this paper, only the results from data taken with the measurement configuration for which the scatter correction parameters exhibit a clear relationship with particle parameters (particle size and concentration) are shown in order to keep the discussion clear and concise. For the 2-component system MSC parameters obtained from the Td spectra and for the 4-component system MSC parameters obtained from the Rd spectra exhibited the clearest relationship with respect to d_p and y . The differences in performance of scatter correction methods in relation to measurement configuration was seen in an earlier simulation study[8] and observations made in this study using experimental data is consistent with that study. Therefore, when applying the method described in this paper for extracting particle size information, the choice of measurement configuration is an important factor.

Initial analysis showed that while EMSC could provide better scatter correction from the point of view of better performing calibration models for particle concentration, for the datasets considered here, the parameters obtained by applying EMSC did not show clear relationship with either d_p or y ,

228 indicating that any information on these properties that may be embedded in the parameters are not
229 easily (if at all) extractable. Therefore MSC which showed clear dependence on particle diameter and
230 concentration is used in the discussions below. It should be noted that it is possible to use EMSC for the
231 step where a calibration model is built to predict the particle concentration $\hat{d}_p$ in order to get better
232 estimates of $\hat{y}$ while using the MSC parameters to obtain the particle diameter information. For sake of
233 simplicity, in this paper we chose MSC for correcting the spectra which is used to build the PLS model
234 for $\hat{y}$ as well as for estimating $\hat{d}_p$ from the MSC parameters.

235

236 3.1 Analysis of scatter correction parameters in two- and four-component systems

237 For the first dataset (polystyrene-water), MSC was applied to the Td spectra after smoothing, and the
238 MSC parameters, a and b , were plotted against d_p and y to investigate the information contained in
239 the parameters. Fig. 2 shows that both parameters vary systematically with the scattering related sample
240 conditions i.e. d_p and y . Figs. 2(a1) and (b1) show the variations in a and b with variations in particle
241 diameter at fixed concentrations. Figs. 2(a2) and (b2) show the variations in a and b with variations in
242 particle concentration at fixed particle diameters. It is clear that the MSC parameters are impacted by
243 both particle concentration and diameter. The variation of both a and b with particle diameter was found
244 to be well explained by a second order polynomial fit for each concentration. This can be seen from the
245 solid curves in Figs. 2(a1) and (b1) which are obtained by regression. The effect of particle
246 concentration on the MSC parameter a at fixed particle diameter required a third order polynomial
247 which is indicated by the solid curves in Fig. 2(a2) while b was found to be well described by a linear fit
248 which is shown by the solid lines in Fig. 2(b2). This analysis suggested the use of equations of the form
249 given by Eqs. (3) and (4). The coefficients in these equations were estimated using least squares
250 regression. The values and 95% confidence intervals of the coefficients in Eqs. (3) and (4) are given in
251 Table 1. The confidence intervals indicate that all the coefficients are significant. Similar analysis was

carried out with Rd spectra of the 4-component system .. Figs. 3(a1) and (b1) show the variations in a and b with variations in particle diameter at fixed concentrations. Figs. 3(a2) and (b2) show the variations in a and b with variations in particle concentration at fixed particle diameters. In this case, second order polynomial curves best described the variations of both a and b with particle diameter at fixed particle concentrations as well as with particle concentration at fixed particle diameters. The solid curves in the subplots of Fig. 3 are the best fit curves obtained by regression in each case. It is observed that, compared to the 2-component system, the MSC parameters for the 4-component system exhibit larger uncertainty in terms of their variations with d_p and y . This leads to higher error in fitting the 4-component samples as can be clearly observed by examining the fitted curves in Fig. 3. This analysis indicates that MSC parameters appear to contain extractable information regarding the scatter-related sample characteristics namely particle size and concentration.

The variations in the MSC parameters at each particle diameter and concentration seen in Fig. 3 suggest that the scatter correction parameters are influenced by one or more factors in addition to particle diameter and concentration. One plausible explanation is that the changes in concentrations of other components in the mixture will result in a change in the refractive index of the suspending medium. This will affect the intensity of light in two ways. It will affect the reflectance/transmittance at the glass boundaries of the cuvette and thus the overall intensity collected by the detector.[12] Also, a change in refractive index of a sample affects the magnitude of light scattered by the particles since light scattering by particles is fundamentally due to the refractive index contrast between the particles and the suspending (liquid) medium.

A simulated dataset consisting of spectra simulated for the same conditions as the samples in the experimental dataset was used to check the above hypothesis. Simulations were based on the Radiative Transfer Theory (RTT) which has been widely used in medical diagnostics and atmospheric sciences to accurately model the propagation of light through turbid media and known to provide good agreement with experimental data [13]. Details of the simulation are given in the supporting information. The

277 absorption and scattering coefficients were calculated by using Mie Theory which accurately models
278 scattering by spherical particles. The bulk absorption coefficients μ_a and the bulk scattering coefficients
279 μ_s obtained using Mie theory are shown in Figures 4(c) and (d), respectively. The effect of change in
280 the refractive index of the mixture due to the change in sample composition is observed from the slight
281 difference between two adjacent μ_s curves in Fig. 4(d). This small difference in the bulk scattering
282 coefficient leads to differences in the spectra of samples which contain the same particle diameter and
283 concentration but different composition of the liquid species in the mixture.

284 In Fig. 5, the relationship between MSC parameters used to correct the simulated Rd spectra (Rd_sim)
285 with concentration and diameter show very similar patterns as observed in Fig. 3 which was obtained
286 from the experimental dataset. The same uncertainty in MSC parameters for samples with the same
287 particle conditions is also observed from the simulated dataset. It should be noted that in the
288 simulations, no instrumental drift or other physical changes that induce variations in the spectra were
289 included. The similarity in the uncertainties in the MSC parameters therefore implies that the
290 baseline/slope change in the spectra of samples with the same sample conditions is due to the difference
291 in refractive index of the samples due to differences in the concentrations of the liquid species which is
292 captured by the MSC method. This conclusion can be made because in the simulations, the refractive
293 index of the suspending medium comprising the liquid species in the mixture is the only physical
294 property that is varying when particle diameter and particle concentrations are fixed. This analysis
295 indicates that the scatter correction parameters are affected not just by particle size and concentration
296 but also to a small extent by the refractive index of the medium. In other words, these parameters are a
297 function of particle diameter, particle concentration and the refractive index of the mixture.

298

299 3.2 Extracting particle size information from scatter correction parameters

300 Given that the particle size information is present in the scatter correction parameters, it would be of
301 interest to know if this information is extractable. Researchers have attempted to obtain particle size

information through applying multivariate calibration models such as PLS to the spectra directly or after correction by empirical preprocessing methods.[15-18] It is however unclear, in these studies, whether it is the particle size or concentration that is modeled since the concentration of the particle in these studies are strongly correlated to the particle size. For instance, Rantanen *et al* reported a method for in-line particle diameter monitoring for high shear granulations in which the particle diameter increases during the process.[18] With the chemical contents in the granulator remaining the same, it implies that the particle number density decreases which can then be related to the changes in the particle diameter. Instead of modeling the particle diameter directly, multivariate regression is likely to model the information related to the particle number density, a correlated factor to the particle diameter, especially on the data preprocessed to remove scatter-related information. Since the effect of particle size on spectra is nonlinear and confounding effects arise due to competing absorption and scattering effects on the spectra, it may be more effective to use the scatter correction parameters. This is because the effect of absorption is decoupled and also because of the possibility of obtaining linear (in the sense of the regression parameters) models relating scatter correction parameters to particle sizes.

In this study, we compared the performance of models for estimating the particle diameter $\hat{d}_p$ using (a) PLS model built on spectra without applying scatter correction ($\mathbf{x}_{meas}$); (b) PLS model built on spectra after applying scatter correction ($\mathbf{x}_{corr}$); and (c) Regression models using MSC parameters and following the methodology described in §2.3. For the approach (c), 3 equations for estimating particle diameter namely, Eqs. (5),(6) and(8) for the 2-component dataset and Eqs. (9)-(11) for the 4 component dataset, were investigated. The two stage approach proposed in §2.3 was tested using cross-validation. The two steps were carried out by using all but one of the samples in stage 1 and applying the resultant model (Stage 2) to the left-out sample. This process is continued till all the samples have been left out from stage 1 once. Table 2 summarizes the performances of the different models for the 2- and 4-component datasets which are discussed in the proceeding sections.

3.2.1 Two-component system

From Table 2 it is seen that using the MSC parameters to estimate $\hat{d}_p$ leads to an appreciable reduction in the estimation errors. Using PLS models built on either $\mathbf{x}_{meas}$ or $\mathbf{x}_{corr}$ leads to similar performance in terms of RMSECV which is also evident in the RMSECV curves for the two models in Fig. 8(a). All the three equations used to predict $\hat{d}_p$ using MSC parameters (Eqs. (5), (6), and (8)) lead to appreciable reduction in the error compared to the PLS models. Eqs. (5) and (6) which use MSC parameters a and b respectively give more or less similar performance with around 55% reduction in error. Eq. (8) which combines the information contained in a and b provides the best performance with around 70% reduction in error. The predicted versus the actual diameters for the two PLS models and the model using Eq. (8) are given in the Supporting Information (Figxx). As mentioned previously the use of Eqs. (5), (6), and (8) for obtaining $\hat{d}_p$ requires the concentration of the particles to be estimated, and this was provided using PLS model built on the spectra for this purpose. Table 2 summarizes the performance of PLS models built on un-corrected $\mathbf{x}_{meas}$ and the scatter-corrected $\mathbf{x}_{corr}$ spectra to predict particle concentration. As expected the estimation error in concentration is lower when $\mathbf{x}_{corr}$ are used. If the scatter correction method is effective in selectively removing the underlying scattering and other non-chemical effect, then it should lead to a better PLS model for predicting particle concentration. . Therefore when using the three equations (Eqs. (5), (6) and (8)), the concentrations of particles estimated from the corrected spectra were provided as input.

3.2.2 Four-component system

In the case of 4-component system, the results were different from that observed in the 2-component dataset. From Table2, the lowest error in predicting particle diameter is obtained using a PLS model built on the spectra without scatter correction ($\mathbf{x}_{meas}$). The PLS model built on $\mathbf{x}_{corr}$ leads to more than 100% increase in the error. .. The best model for predicting the particle diameter using the MSC parameters was given by Eq. (11) which combines information in a , b , and $\mathbf{x}_{corr}$. Unlike the 2-component system, the error in this case is more than 100% higher compared to the PLS model using

351 $\mathbf{x}_{meas}$. The reason for this was investigated first by examining the performance of the PLS model to
 352 predict particle concentration which is an input for Eq. (11). From Table 2, it is seen that RMSECV for
 353 the estimated concentration is much higher compared to the 2-component dataset. Both $\mathbf{x}_{meas}$ and $\mathbf{x}_{corr}$
 354 give similar levels of error in the estimated concentration though the model built on $\mathbf{x}_{corr}$ requires fewer
 355 numbers of latent variables. If the large error in estimated diameter $\hat{d}_p$ is due to the error contributed
 356 by $\hat{y}$, then by replacing $\hat{y}$ by the actual concentration y should result in significant improvement and
 357 lead to similar performances that seen for the 2-component dataset. However, the error in estimated $\hat{d}_p$
 358 did not reduce significantly indicating that the source of this increase in error lies elsewhere.

359 Further investigation was carried out by examining the concentrations of the different components and
 360 their correlation structure. The 4-component dataset was designed to eliminate the concentration
 361 correlation between the polystyrene particles and other components of the system. However, in the
 362 dataset the particle diameter is weakly correlated to the main constituents of the medium, H₂O and D₂O
 363 with a correlation coefficient of about 0.26 with each of these components. This raises the possibility
 364 that the PLS model built on $\mathbf{x}_{meas}$ for estimating particle diameter will be improved by such a
 365 correlation. Examining the scores of the PLS model, it was found that the scores of the first latent
 366 variable and to a certain extent the second latent variable are linearly related to d_p , as indicated in Figs.
 367 12(b1) and (b2). Examining the loadings of these two latent variables shown in Figs. 12(a1) and (a2),
 368 we see that they appear to be explaining variations that affect the baseline of the spectra i.e. light
 369 scattering. Applying MSC and then building a PLS model on $\mathbf{x}_{corr}$ would result in the removal of
 370 information regarding particle diameter and should lead to models with higher errors in the estimation
 371 of particle size. The scores of the first and second latent variables obtained by applying PLS to $\mathbf{x}_{corr}$ in
 372 Figs. 13(b1) and (b2) shows that the first latent variable no longer possesses a clear relationship with
 373 particle diameter. Also the first latent variable now resembles more like the second LV for the un-
 374 corrected spectra (Fig. 12(a2)). However, there is no significant pattern in this case with respect to d_p .

375 It is also interesting to note that the number of latent variables required for the PLS model to predict
376 particle diameter is reduced from 7 when $\mathbf{x}_{meas}$ is used to 4 when $\mathbf{x}_{corr}$ is used. This explains the increase
377 in the error in the estimated particle diameter when PLS is applied after scatter correction. Despite this
378 removal of particle size information, the model obtained from $\mathbf{x}_{corr}$ is still statistically significant and
379 almost of similar level of performance as the models using the scatter correction parameters to estimate
380 particle size. This is probably due to the fact that $\mathbf{x}_{corr}$ still has chemical information regarding H₂O and
381 D₂O which are in turn correlated to the particle diameter thus providing the ability to predict particle
382 diameter despite most of the information regarding this parameter has been removed by scatter
383 correction. The MSC parameters on the other hand, do not include the correlation between particle size
384 and the concentrations of H₂O and D₂O, since these parameters are indicative of baseline and slope
385 changes in the spectra while absorptivity changes (and thus information) due to concentration changes
386 in H₂O and D₂O remain in the corrected spectra.

387 Recalling that the MSC parameters for the 4-component dataset are affected by particle size,
388 concentration and the refractive index of the suspending medium (§3.1), it should be pointed out that the
389 models relating particle diameter to the MSC parameters were developed by neglecting the effect of the
390 refractive index changes. This could also potentially lead to an increase in the error in estimating
391 particle diameter. A further point to be noted is that for the 4-component system, the prediction of
392 particle size by using equations that arise from inverting the expressions relating a or b (i.e. Eqs. 9) and
393 (10)) led to two positive values for the particle diameter when the quadratic equations are solved. The
394 ambiguity resulting from this meant that the expressions were not practically usable and therefore the
395 results pertaining to these inverted equations are not shown in Table 1. This problem was not
396 encountered when the combined Eq. (11) was used. Since the equations relating the MSC parameters to
397 particle diameter and concentration that are given here are not necessarily unique, it may be possible to
398 develop an alternative regression model to overcome this problem.

399 4. Conclusions

400 This study provides an insight into the nature of information contained in the scatter correction
401 parameters. It shows that a scatter correction technique which leads to better calibration models for
402 estimating concentration of chemical species need not necessarily be the best in terms of the scatter
403 correction parameters containing extractable information. It was found that the MSC parameters
404 contained significant information regarding scatter-causing properties namely particle size and
405 concentration. The parameters from EMSC which leads to better performing calibration models
406 compared to MSC do not show a clear relationship with the scatter-causing properties. This may be due
407 to the fact that the information is spread over a larger number of parameters and also the possibility that
408 EMSC might be removing other non-chemical variations that may be presented in the dataset. Further,
409 whether a clear relationship between the MSC parameters and the particle size and concentration was
410 observed depended strongly on the measurement configuration, indicating that the performance of a
411 scatter correction technique will depend on the measurement configuration. This is in line with the
412 observations made in an earlier study based on simulations.[8]

413 Given that the information regarding particle size is present in the MSC parameters, a method to
414 extract this information was proposed and evaluated using the two-component and four-component
415 datasets. It was found that for the 2-component dataset, the method was effective in extracting this
416 information and the model resulting from this method led to a reduction of about 70% in the error in the
417 estimation of particle size compared to models obtained by applying PLS to the spectra. For the 4-
418 component dataset, the error in using the proposed method was considerably higher. This appears to be
419 due to the increased uncertainty contributed by the changes in the refractive index of the suspending
420 medium which is not included in the model. Also the PLS model built on the spectra led to considerably
421 lower error compared to the proposed method. Analysis indicates that this is due to chance correlations
422 between particle diameter and the concentrations of D₂O and H₂O present in the mixture.

423

424 REFERENCE

- 425 [1] A. Rinnan, F.v.d. Berg, S.B. Engelsen, Trends Anal. Chem. 28 (2009) 1201.
426 [2] P. Geladi, D. MacDougall, H. Martens, Appl. Spectrosc. 39 (1985) 491.
427 [3] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Appl. Spectrosc. 43 (1989) 772.
428 [4] S. Wold, H. Antti, F. Lindgren, J. Öhman, Chemom. Intell. Lab. Sys. 44 (1998) 175.
429 [5] H. Martens, J.P. Nielsen, S.B. Engelsen, Anal. Chem. 75 (2003) 394.
430 [6] Z.-P. Chen, J. Morris, E. Martin, Anal. Chem. 78 (2006) 7674.
431 [7] K. Wang, G. Chi, R. Lau, T. Chen, Anal. Lett. 44 (2011) 824.
432 [8] S.N. Thennadil, E.B. Martin, J. Chemom. 19 (2005) 77.
433 [9] R. Steponavicius, S.N. Thennadil, Anal. Chem. 81 (2009) 7713.
434 [10] R. Steponavicius, S.N. Thennadil, Anal. Chem. 83 (2011) 1931.
435 [11] S.N. Thennadil, H. Martens, A. Kohler, Appl. Spectrosc. 60 (2006) 315.
436 [12] C.F. Bohren, D.R. Huffman, Absorption and Scattering of Light by Small Particles, Wiley-VCH,
437 Berlin, 2004.
438 [13] A. Engdahl, B. Nelander, J. Appl. Phys. 86 (1987) 1819.
439 [14] J.P. Devlin, J. Appl. Phys. 112 (2000) 5527.
440 [15] M.M. Reis, P.H.H. Araújo, C. Sayer, R. Giudici, Macromol. Rapid Commun. 24 (2003) 620.
441 [16] K. Ito, T. Kato, T. Ona, J. Raman Spectrosc. 33 (2002) 466.
442 [17] A. Gupta, G.E. Peck, R.W. Miller, K.R. Morris, J. Pharm. Sci. 93 (2004) 1047.
443 [18] J. Rantanen, H.k. Wikström, R. Turner, L.S. Taylor, Anal. Chem. 77 (2004) 556.
444
445

Figure 1

Fig. 1. A flowchart of the methodology used for estimating particle diameter and concentration. The method involves Stage 1 : Calibration model building (steps in black) and Stage 2 : Prediction of particle diameters and concentrations of unknown samples (steps in blue).

Figure 2

Fig. 2. (a) Measured total transmittance spectra (x_{meas}) of polystyrene-water 2-component dataset. (b) MSC preprocessed spectra (x_{corr}) using the mean of x_{meas} as a reference spectrum.

Figure 3

Fig. 3. (a) Changes in MSC parameter a in the 2-component system with (a1) particle diameter and (a2) concentrations. (b) Changes in MSC b with (b1) particle diameter and (b2) concentrations. Solid curves were generated from the best fit obtained using least squares regression.

Figure 4

Fig. 4. (a) Measured total reflectance spectra (x_{meas}) of the 4-component dataset. (b) MSC preprocessed spectra (x_{corr}) using the mean of x_{meas} as a reference spectrum.

Figure 5

Fig. 5. (a) Changes in MSC parameter a in the 4-component system with (a1) particle diameter and (a2) concentrations. (b) Changes in MSC parameter b with (b1) particle diameter and (b2) concentration. Solid curves were generated from the best fit obtained by least squares regression.

Figure 6

Fig. 6. (a) Simulated total reflectance spectra (Rd_{sim}) of the 4-component dataset. (b) MSC preprocessed spectra ($Rd_{sim_{corr}}$) using the mean of Rd_{sim} as a reference spectrum. The bulk absorption and scattering coefficients used for the simulation are in (c) and (d), respectively.

Figure 7

Fig. 7. Results of simulated spectra (Rd_sim) of the 4-component system after MSC preprocessing. (a) Changes in MSC parameters a with (a1) particle diameter and (a2) concentrations. (b) Changes in MSC b with (b1) particle diameter and (b2) concentration. Solid curves were generated from the best fit obtained by least squares regression.

Figure 8

Fig. 8. (a) RMSECV curves of PLS models for estimating particle diameter in 2-component system from x_{meas} and x_{corr} . (b1) and (b2) are the predictions using PLS models built on x_{meas} and x_{corr} , respectively. (b3) is estimated using the inversion Eq. (A.8) in Supplementary Information which combines MSC parameters and x_{corr} .

Figure 9

Fig. 9. (a) RMSECV curves of PLS models for estimating particle concentration in the two-component system. (b1) and (b2) show plots of estimated versus actual values of particle concentration in the system for PLS models built on x_{meas} and x_{corr} , respectively.

Figure 10

Fig. 10. (a) RMSECV curves of PLS models for estimating particle diameter in 4-component system. (b1) and (b2) are the prediction using PLS models built on x_{meas} and x_{corr} , respectively. (b3) is estimated using inversion Eq. (A.11) in Supplementary Information which combines MSC parameters and x_{corr} .

Figure 11

Fig. 11. (a)RMSECV curves of PLS models for estimating particle concentration in the four-component system. (b1) and (b2) show plots of estimated versus actual values of particle concentration in the system for PLS models built on x_{meas} and x_{corr} , respectively.

Figure 12

Fig. 12. (a1)-(a3) loading curves and (b1)-(b3) scores of the first 3 loadings of the PLS model built on x_{meas} to estimate particle diameter.

Figure 13

Fig. 13. (a1)-(a3) loading curves and (b1)-(b3) scores of the first 3 loadings of the PLS model built on x_{corr} to estimate particle diameter.