Extraction of Chemical Information of Suspensions using Radiative Transfer Theory to Remove Multiple Scattering Effects: Application to a model two-Component System

Raimundas Steponavičius[1], *Suresh N. Thennadil*[1,2]

[1]Merz Court, School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, NE1 7RU, United Kingdom; [2] James Weir Building, 75 Montrose Street, Department of Chemical and Process Engineering, University of Strathclyde, Glasgow, G1 1XJ, United Kingdom.

*To whom correspondence should be addressed: suresh.thennadil@strath.ac.uk.

RECEIVED DATE (to be automatically inserted after your manuscript is accepted if required according to the journal that you are submitting your paper to)

ABSTRACT

An approach for removing multiple light scattering effects using the radiative transfer theory (RTE) in order to improve the performance of multivariate calibration models is proposed. This approach is then applied to the problem of building calibration models for predicting the concentration of a scattering (particulate) component. Application of this approach to a simulated four component system showed

that it will lead to calibration models which perform appreciably better than when empirically scatter corrected measurements of diffuse transmittance (T_d) or reflectance (R_d) are used. The validity of the method was also tested experimentally using a two-component (Polystyrene-water) system. While the proposed method led to a model that performed better than that built using R_d, its performance was worse compared to when T_d measurements were used. Analysis indicates that this is because the model built using T_d benefits from the strong secondary correlation between particle concentration and pathlength travelled by the photons which occurs due to the system containing only two components. On the other hand, the model arising from the proposed methodology uses essentially only the chemical (polystyrene) signal. Thus this approach can be expected to work better in multi-component systems where the pathlength correlation would not exist.

KEYWORDS: Scatter correction, Multivariate calibration, Near-infrared spectroscopy, Multiple scattering, Radiative transfer equation, Adding-doubling method.

INTRODUCTION

Accurate estimation of concentrations of chemical components in turbid samples using spectroscopic techniques is still an open-end problem that challenges chemometricians and other applied scientists.[1] The effective solution to this problem is of tremendous practical importance since it is encountered in many areas such as monitoring polymerization[2-5] and fermentation[6,7] processes and in medical diagnostics[8]. Spectroscopic measurements have the potential for providing such information and would be preferable, because they are fast, cheap, compatible with fibre optics and multifunctional.

The main problem, in the quantitative analysis of turbid samples using Near-infrared (NIR) spectroscopy, is that multivariate calibration models built on conventional spectroscopic measurements such as transmittance or reflectance are adversely affected by variations arising from multiple light scattering, because these variations are not necessarily related to changes in chemical information i.e. concentrations of chemical components. There are essentially two ways to deal with undesirable

scattering effects in NIR measurements: remove/minimize them by means of empirical pre-processing or separate scattering effects from absorption by invoking light propagation theory such as the radiative transfer theory. In either case, the goal is to obtain a measure of absorption per unit length, which is independent from variations in pathlength of photons that occur due to multiple scattering and linearly proportional to concentrations of constituents.

The most frequently used approach for scatter correction has been empirical pre-processing because of its simplicity. A number of techniques such as standard normal variate, orthogonal signal correction, multiplicative scatter correction (MSC) and extended MSC (EMSC) have been used to reduce multiple scattering effects[9]. Although these may be sufficient when the variations due to scattering are small and when the signal of the target analytes is large, they are not adequate when variations in scattering are high as is usually the case in many industrial situations. Recently, a physics-based EMSC has been proposed, where the physics of light transport is incorporated to further improve the removal of scattering effects[10].

Very little work has been done on the applicability of the second approach for scatter correction in process analytics mainly due to complex measurements and theory required to extract the absorption and scattering properties of a sample. The measurement techniques for deconvolution of scattering and absorption properties were developed primarily for biomedical applications. The only application of this scatter correction approach for development of Process Analytical Technologies (PAT) was reported by Abrahamsson et. al.[11] They applied it on pharmaceutical tablets and showed a significant improvement in the accuracy of predictions in comparison with direct application of Partial Least Squares (PLS) regression on transmittance measurements. The methodology they used has several shortcomings: two instruments are required (time resolved and conventional spectrometer), diffusion approximation assumptions have to be met and the reduced scattering coefficient could not be measured beyond 1100 nm using their time resolved system (whereas the overtones of the organic compounds appear in the NIR region above 1100 nm) and therefore it had to be extrapolated for the higher wavelength region.

The aim of this research was to develop a methodology for estimation of chemical information in suspensions where the radiative transfer theory is used to remove multiple scattering effects. Broadly, the problem of extracting concentration of chemical species in a particulate system (suspensions or powder mixtures) can be classified into two groups viz. the extraction of information of a chemical species that (a) purely absorbs or (b) both absorbs and scatters light.

The work reported in this paper consists of two parts: simulation and experiment. Simulation study was used to show the maximum theoretical improvement in the prediction accuracy possible using the methodology described in this paper. A simple polystyrene-water system was used to evaluate the performance of the proposed methodology on an experimental dataset. This paper focuses on the problem arising when the chemical species of interest both absorbs and scatters light.

THEORY

A turbid sample is a heterogeneous sample that scatters light e.g. particles suspended in water. The problem in obtaining a good calibration model for turbid samples using conventional chemometrics stems from the fact that the measured change in absorbance or optical density cannot be effectively correlated with changes in concentrations of chemical components because it is confounded with changes caused by light scattering. The problem is illustrated in figure 1.

Figure 1(a) represents case (a) where light passes through a homogenous liquid mixture. In this case, the photons only undergo absorption. Figure 1(b) represents case (b) where light passes through a turbid sample (particles suspended in liquid) with particle concentrations sufficiently low that the photons passing through the sample encounter a particle only once (single scattering). Figure 1(c) represents case (c) where the particle concentration is sufficiently high such that the photons encounter several particles i.e. undergo multiple scattering events before exiting the sample. Since the direction of the photons change during each scattering event, the total pathlength travelled by the photon before exiting the sample will be different from (greater than or equal to) the sample thickness. From the point of making measurements for estimating the concentrations of chemical components, it is desirable to

choose a configuration that provides measurements that are proportional to the concentrations of the chemical components. For case (a) this is achieved in a straightforward manner by measuring the axially (collimated) transmitted light and applying Beer-Lambert's law which for a sample containing n chemical species is given by:

$$A(\lambda) = -\ln(T_c) = -\ln\left(\frac{I(\lambda)}{I_0(\lambda)}\right) = \ell \cdot \sum_{i=1}^{n} \sigma_{a,i}(\lambda) \cdot c_i \qquad (1)$$

where $\sigma_{a,i}$ is the absorption cross section of chemical species i, c_i is its concentration and λ is the wavelength of light. The pathlength travelled by the photons in this case is the sample thickness ℓ. In case (b) such a measurement can still be made, though the Beer-Lambert equation needs to be modified as:

$$A(\lambda) = -\ln(T_c) = -\ln\left(\frac{I(\lambda)}{I_0(\lambda)}\right) = \ell \cdot \sum_{i=1}^{n} \sigma_{ext,i}(\lambda) \cdot c_i \qquad (2)$$

where $\sigma_{ext,i} = \sigma_{a,i} + \sigma_{s,i}$, is the extinction cross-section and $\sigma_{s,i}$ is the scattering cross-section which is non-zero for those species which are particles. The scattering cross-section is a highly non-linear function of particle size and shape. In this case, even though Beer's law applies, the situation is complicated by the presence of non-linear scattering effects since for the same concentration of the scattering species, two different particle sizes would lead to changes in the absorbance which need to be corrected when building calibration models. In Case (c), the extent of multiple scattering becomes too high which precludes accurate measurement of the un-scattered axially transmitted light both due to the small fraction of light that would have managed to travel without being scattered as well as due to the fact that as the amount of scattering increases, an increasing amount of forward scattered light will be included in the measurement. Thus for such turbid samples, either diffuse reflectance or diffuse transmittance measurements are made. Since these measurements (schematically shown in figure 1(b) and (c)) involve collection of light exiting from the sample in all directions, the average pathlength travelled by the photons is no longer equal to the sample thickness and it is not constant from sample to

sample with the variation depending on the variation in the scattering properties of the sample which in turn depends on the particle size, shape and concentration.

In the first case, multivariate calibration is straight-forward because absorption varies only with concentrations of chemical components (the other two terms, the absorption cross section and the path length are constant) and this relationship is linear (see eq. 1). Therefore PLS models that are based on the assumption of linear relationship between the absorbance and concentrations of the species, usually give very good results. In the second case, however, two terms in equation 2 can vary: the concentrations and the extinction cross-section of particles (the pathlength of light is constant). In the third case, the concentrations, the extinction cross-section of particles and the pathlength of light can all vary. This leads to confounding effects in the estimation of concentration of chemical components in turbid samples which arise because different combination of values of concentration, pathlength and extinction cross-section can lead to the same measurement value $A(\lambda)$ in equation (2). From the point of inverting the measurement value to obtain the concentration of species, since the contribution from these 3 parameters due to changes in particle size, shape and concentration cannot be distinguished from each other, it could cause potentially large errors in the estimated concentrations and thus will result in lack of robustness. In addition, these variations are nonlinear and therefore they degrade linear calibration models. Thus to obtain accurate calibration models for turbid samples, variations in the pathlength and the absorption cross section of particles have to be corrected.

The main source of variation in absorption in the multiple scattering regime is the pathlength of photons, see fig. 1 c.). In principle, this variation can be eliminated by obtaining a measure of absorption per unit length, which is independent of pathlength, using radiative transfer theory. The change in the intensity of light of a given wavelength travelling through a sample in a certain direction is described by the radiative transfer equation (RTE):[12]

$$\frac{d\boldsymbol{I}(\boldsymbol{r},\boldsymbol{s})}{d\boldsymbol{s}} = -\mu_t \cdot \boldsymbol{I}(\boldsymbol{r},\boldsymbol{s}) + \frac{\mu_s}{4\cdot\pi}\int_{4\pi} p(\boldsymbol{s},\hat{\boldsymbol{s}})\cdot \boldsymbol{I}(\boldsymbol{r},\hat{\boldsymbol{s}})\cdot d\omega \qquad (3)$$

where $I(r,\boldsymbol{s})$ is the specific intensity at a distance r from source along directional vector $\boldsymbol{s}$, μ_a (cm^{-1}) is the bulk absorption coefficient, μ_s (cm^{-1}) is the bulk scattering coefficient, $\mu_t = \mu_a + \mu_s$ is the total extinction coefficient, $p(\boldsymbol{s},\hat{\boldsymbol{s}})$ is the phase function, which is a measure of the angular distribution of scattered light and ω is the solid angle. The bulk absorption and scattering coefficients are proportional to concentrations of absorbing and scattering components respectively. For a system with multiple components the bulk absorption and scattering coefficients are the sum of the respective coefficients of individual components:

$$\mu_a = \sum_{i=1}^{n} \mu_{a,i} = \sum_{j=1}^{n_p} \sigma_{ap,j} \cdot c_{p,j} + \sum_{k=1}^{n_a} \sigma_{a,k} \cdot c_k \tag{4}$$

$$\mu_s = \sum_{i=1}^{n} \mu_{s,i} = \sum_{j=1}^{n_p} \sigma_{sp,j} \cdot c_{p,j} \tag{5}$$

where $\sigma_{ap,j}$and $\sigma_{sp,j}$ are the absorption and scattering cross-sections (cm^2) of the particulate species j, $c_{p,j}$ is the concentration of the particulate species j expressed as number density i.e. number of particles per unit volume (cm^{-3}) and n_p is the number of different particulate species present in the sample. $\sigma_{a,k}$ represents the absorptivity (cm^2/g) of the purely absorbing species k, c_k is the concentration (g/cm^3) of the absorbing species $k$and n_a is the number of purely absorbing species present in the sample. It should be noted that the bulk absorption and scattering coefficients as well as the absorption and scattering cross-sections of the particles and the absorptivity of the purely absorbing species are all wavelength dependent. In equation 4 μ_a has been split into two terms. The first summation represents the contribution from the particulate species and the second summation represents the contribution from the purely absorbing species. The first term will vary both due to the concentration of the particulate species as well as its particle size because the absorption cross-section $\sigma_{ap,j}$ is dependent on the particle size and shape. The second term varies only with the concentration of the purely absorbing species. In equation 5 the contribution to the scattering coefficient is only from the particulate species present in the sample. The scattering cross-sections of the particulate species are dependent on particle size and shape.

The phase function p describes the angular distribution of scattered light at a particular wavelength. There are several phase functions that have been used among which the most common is the Henyey-Greenstein phase function:

$$p(\theta,g)=\frac{1-g^2}{\sqrt{\left(1+g^2-2\cdot g\cdot\cos\theta\right)^3}} \tag{6}$$

where $\theta-$ is an angle between incident and scattered directions and g is the anisotropy factor. As we can see from equations 3-6, at each wavelength, the RTE is defined by three variables μ_a, μ_s and g (called bulk optical properties). To extract them the inverse RTE has to be solved. Since there are three parameters involved, to invert the RTE we need atleast three measurements at each wavelength. One can notice that μ_a is a measure of absorption per unit length (cm^{-1}) and it is independent of the pathlength travelled by the photons. However, it is still not free from scattering variations due to changes in absorption cross section of particles σ_{ap}, which is a function of particle size and shape.

MATERIALS AND METHODS

RTE based scatter correction and calibration approach

The proposed methodology for estimation of concentrations of chemical components in suspensions is outlined in figure 2. Essentially, it is a two step procedure: acquisition of the bulk optical properties and then extraction of pertinent chemical information from μ_a. The other two properties μ_s and g can be potentially used to correct μ_a for changes in $\mu_{a,p}$ though such approaches are not considered in this paper. As mentioned earlier, in order to extract the bulk optical properties by inverting the RTE, atleast three measurements are required. The three measurements used in this study for extraction of the optical properties were: total diffuse transmittance T_d, total diffuse reflectance R_d and collimated transmittance T_c. To obtain the bulk optical properties from these measurements we need to invert the radiative transfer equation. There is no analytical solution to the radiative transfer equation, but there are a few numerical solutions such as Adding-Doubling (AD)[13], Monte Carlo (MC)[14] and Discrete Ordinates[12] method. Of these the AD and the MC methods are the most frequently used. The MC method can

accommodate any type of measurement geometry (e.g. spatially-resolved measurements which measure reflectances at specific distances from the incident beam) and can also take into account incident beam shape (collimated, diverging etc.), finite beam width and sample width[14]. The disadvantage is that this approach is computationally very intensive. The AD method is much faster but does not take into account beam width and assumes that the sample is of infinite width thus ignoring any light loss through the sides of the sample which in some cases could lead to significant errors. This method is well suited for computing total diffuse reflectance and transmittance measurements but cannot be used for computing spatially-resolved measurements. The AD method has been widely used to extract optical properties from chemical and biological systems (bacteria, blood, tissue etc.) when integrating sphere measurement set-ups are used to measure total diffuse reflectance and transmittance measurements[13,15,16] since computationally it is much faster than the MC method. Therefore, in this work, the inverse Adding-Doubling algorithm (IAD) which iteratively applies the AD method was used to invert the RTE[13].

The influence of particle size changes on the accuracy of predictions was also investigated in this work. After extracting the bulk absorption coefficient using the RTE to invert the measurements, the resulting absorption spectra (i.e. the bulk absorption coefficient as a function of wavelength) is used for building a multivariate calibration model for estimating the concentration of the chemical component of interest. Before building the models additional empirical pre-processing to further reduce unwanted variations could also be included as part of this approach.

SIMULATION

The proposed approach was applied to a simulated dataset of spectra of turbid samples (considered previously by Thennadil and Martin[9]) to test the extent of improvement in model performance that can be theoretically obtained compared to the performance that would be obtained using empirical scatter-correction approaches. In their work, Thennadil and Martin modelled the turbid system as a four component system comprising one scattering component (polystyrene particles) and three non-

scattering components which were simulated using the optical properties of toluene (species 2), deuterated water (species 3) and water (species 4). The volume fraction of particles varied between 0.01 and 0.1 and the radius of particles spanned the range 100nm to 500nm. The volume fraction of species 2 and 3 spanned the ranges 0-0.0115 and 0.2-0.4 respectively. The spectra were simulated using the radiative transfer equation. Noise was then added to the spectra to resemble the real measurements. Thennadil and Martin used the dataset so created to study the effectiveness of various pre-processing techniques on calibration models built for predicting the concentration of a non-scattering component. In the current study this dataset consisting of 50 calibration samples and 391 test set samples was used to compare the proposed approach of using the extracted bulk absorption spectra for building calibration models for a scattering component with those obtained using the traditional approach of using diffuse reflectance or transmittance measurements with empirical pre-processing to remove scattering effects.

EXPERIMENT

Design of experiments. To test the methodology on experimental data a simple two component system, polystyrene particles suspended in deionised water was used with the aim of estimating the concentration of polystyrene particles from NIR measurements. The samples were prepared according to the following experimental design – five particle sizes: 100nm, 200nm, 300nm, 430nm and 500nm, seven concentrations (in wt.%) for each particle size: 0.1%, 0.5%, 0.9%, 1.23%, 1.6%, 1.95 and 2.3%, giving a total of 35 samples.

It should be noted that there are two major differences between the system studied using simulations and the model system used to generate the experimental dataset. The first is that in the simulation a 4 component system was considered whereas the experimental model system has only 2 components. Secondly, the simulation dataset spans a much larger range of concentrations of scattering component (1-10%). Thus the highest particle concentration in the simulated dataset is almost 5 times larger than that considered in the model experimental system. The implication of the latter point is that multiple scattering effects would be much more dominant in the simulated system. Since the maximum particle

concentration in the experimental system is only 2.3%, the multiple scattering effects will be comparatively small. As a result, in this regime (two component system with low multiple scattering) we would expect calibration models based on single measurements (reflectance or transmittance) with empirical scatter correction approaches to work reasonably well. Such a relatively simple system was chosen since it would allow us to examine the accuracy of the complex inversion steps and the instrumentation setup involved in the extraction of the bulk optical properties. If the bulk absorption coefficient is extracted with sufficient accuracy, then the calibration model built using this approach will perform as well as or better than the single measurement approaches. This would validate the approach in terms of its accuracy as well as highlight any problems in the inversion methodology that need to be addressed for the successful implementation of this method for more complex systems.

Measurement setup. Three spectroscopic measurements $\boldsymbol{T}_{\mathbf{c}}$, $\boldsymbol{T}_{\mathbf{d}}$ and $\boldsymbol{R}_{\mathbf{d}}$[*] were taken for each sample using a scanning spectrophotometer (CARY 5000, Varian Inc.) with a diffuse reflectance accessory. Spectral data was collected in the wavelength region 1600-1848 nm at 4 nm intervals resulting in measurements at 63 discrete wavelengths per spectrum. This region was chosen because the first overtone peaks of polystyrene due to C-H stretching vibrations appear around 1680 nm. The Peltier (TE) cooled PbS detector was used for this wavelength region. An average integration time was set to 0.4 s, the bandwidth and the energy level were automatically adjusted to obtain good signal-to-noise ratio. The collimated transmittance ($\boldsymbol{T}_{\mathbf{c}}$) was measured with the instrument's standard configuration. For the total diffuse reflectance ($\boldsymbol{R}_{\boldsymbol{d}}$) and total diffuse transmittance ($\boldsymbol{T}_{\boldsymbol{d}}$) measurements an external diffuse reflectance accessory (DRA-2500, Varian Inc.) was mounted. It consists of a 150 mm diameter integrating sphere (Labsphere), which has a port-to-sphere area ratio of less than 10%. The sphere is coated with "Spectralon" material, which acts as an almost perfect Lambertian surface. A schematic representation of the different measurement configurations is shown in figure 1. In the case of a total

[*] The bold symbols are used to indicate that these are vectors of values over a wavelength range to differentiate from values at single wavelengths represented by the same symbols.

diffuse transmittance measurement, the sample is placed at the entrance port of the sphere, while the exit port is blocked with "Spectralon" reflectance standard. In this way, both collimated and diffusely transmitted light is collected by the detector. For a total reflectance measurement, the sample is placed at the exit port of the sphere, so that all light reflected by the sample is collected in the sphere. To obtain similar irradiation conditions for transmittance and reflectance measurements (i.e. illumination area and angle) different focusing lenses were used.

Extraction of the optical properties

The measurement set-up and the algorithm used for the extraction of the optical properties is similar to the ones used in previous studies[15-17].The extraction of the optical properties was carried out by an iterative method to invert the RTE with the inversion being carried out at one wavelength at a time. For each wavelength, initial guess values of the bulk optical properties μ_a, μ_s and g are given as input. These are used to compute the albedo $a = \mu_s / (\mu_a + \mu_s)$ and optical depth (turbidity) $\tau = (\mu_a + \mu_s) \cdot l$. This is because the adding-doubling equations are written in terms of these parameters and g. For the given parameters the RTE is solved using the adding-doubling method to obtain the calculated values of total diffuse reflectance and total diffuse transmittance which also takes into account the boundary effects (cuvette-sample, cuvette-air boundaries) through the use of Fresnel equations. The calculated values are compared with the experimental values of the three measurements and the process is repeated by suitably updating the guess-values of the parameters until convergence is reached. This inversion step was carried out using nonlinear constrained optimization (MATLAB® 'fmincon' optimizer). The length of the error vector was chosen as the objective function to minimize:

$$f = \sqrt{\left(T_d - \hat{T}_d\right)^2 + \left(R_d - \hat{R}_d\right)^2} \qquad (7)$$

$\hat{T}_d$ and $\hat{R}_d$ are estimates of total diffuse transmittance and reflectance at a specific wavelength. In previous work where the adding-doubling method was used to estimate the bulk optical properties[15,16], the objective function for minimization had an extra term $(T_c - \hat{T}_c)^2$ within the square root in (7) i.e. it

also utilized the error in the calculated value of T_c compared to the experimental values. In the present study, it was found that instead of directly using collimated transmittance in the objective function, the inversion was more stable if the measured optical depth τ (which is just $-\ln(T_c)$) was used as a constraint. If τ was measured accurately we could use it as an equality (hard) constraint which would speed up the optimization. However, at higher turbidity the mismatch between the measured and the actual τ became significant due to the fact that the light collected in the collimated transmittance mode begins to be "contaminated" with light that has undergone scattering mainly due to the amount of forward scattered light becoming significant at higher concentrations. In this case a soft constraint (upper bound < τ < lower bound) works better. It was found that for the lowest concentrations 0.1% and 0.5% by weight, the equality constraint was adequate because the theoretical turbidity matched well with the measured turbidity. At higher concentrations, because of the increasing mismatch, soft constraints were used.

The other inputs/constants to the adding-doubling routine are: the pathlength of cuvette and refractive indices of air, cuvette glass and sample. The refractive indices are needed to compute reflections at the interfaces using Fresnel equations. A 1 mm pathlength cuvette made out of special optical glass (100.099-OS, Hellma) was used in the study. The refractive index of the cuvette for the required wavelength region was provided by the manufacturer (Hellma). The refractive index of polystyrene was taken from Velazco-Roa and Thennadil.[17] However, the values were available only up to 1400 nm. Therefore, the values at higher wavelengths were obtained by extrapolation using the model given by the Cauchy dispersion formula[18]:

$$n(\lambda) = A + \frac{B}{\lambda^2} + \frac{C}{\lambda^4} \tag{8}$$

The refractive index of water was taken from Segelstein[19] and the refractive index of the sample was calculated as a sum of refractive indexes of water and polystyrene multiplied by their respective weight fractions. The refractive index of air was taken as equal to one across the entire wavelength region.

Calibration. Multivariate calibration was carried out using PLS regression. For the simulated data, the same model building and latent variable selection procedures used in [9] were carried out. For the experimental data, accuracy of predictions was evaluated using root mean square error of cross validation (RMSECV). Cross validation was carried out using the 'leave-one-out' method. Further, for the experimental data, all the raw spectra were smoothed using a Savitsky-Golay filter with window width 9 and order 3 to remove noise in the measurements. The computations were carried out using Matlab® and the PLS models were built using PLS_Toolbox by Eigenvectors Research Inc.

RESULTS AND DISCUSSION

SIMULATION

The simulated spectra of total diffuse transmittance and total diffuse reflectance and the corresponding bulk absorption coefficient for the turbid system toluene-polystyrene-water-heavy water are shown in figure 3. All graphs are provided in the same scale so that the magnitudes of variation in each case can be visually compared. It can be seen that (baseline) variation in $\boldsymbol{\mu_a}$ is much smaller than in the two direct measurements: approximately five times smaller than in diffuse transmittance and four times smaller than in diffuse reflectance. Variation in $\boldsymbol{T_d}$ and $\boldsymbol{R_d}$ measurements is due to changes in both chemical and physical properties of the sample, whereas variation in $\boldsymbol{\mu_a}$ is predominantly due to changes in chemical information (concentrations). It is apparent from this comparison that the variation in the pathlength travelled by photons, which is subject to number, size and shape of particles is the main contributor to the variation in spectroscopic measurements of a turbid sample. Not only is the magnitude of variation due to changes in physical properties much larger than due to changes in concentrations, but it is also nonlinear, which makes it a serious problem for the linear multivariate calibration.

Table 1 summarizes the results of the PLS calibration performances using reflectance and transmittance data compared with results obtained when the bulk absorption coefficients $\boldsymbol{\mu_a}$ are used to predict the concentration of the scattering species (polystyrene) in the simulated 4 component system. For the case where diffuse reflectance spectra were used for building calibrations, it was found that pre-

processing by any of the techniques considered in [9] did not improve performance. The EMSCL (Extended Multiplicative Scatter Correction with wavelength dependent log term) method which was found in [9] to be the best performing scatter correction technique for predicting the concentrations of non-scattering species, in the present case, needed fewer number of latent variables (LVs) and therefore is reported. When diffuse transmission measurements were used pre-processing with EMSCL provided a slight improvement.

From Table 1, it is seen PLS models built on μ_a for estimation of concentration of scattering component yielded much better prediction results than those built on diffuse reflectance spectra or diffuse transmittance spectra, the latter exhibiting the worst performance. The RMSEP value obtained by using μ_a was more than 1.7 times lower than that obtained using diffuse reflectance and was achieved with half the number of LVs.

There are two points to be noted in Table 1. Firstly, even for the theoretical situation, the prediction error in particle concentration when using μ_a, while better than using reflectance spectra coupled with pre-processing, is appreciable. Secondly, since there are only 4 components in the system, taking closure condition into consideration, if scattering effects were completely eliminated, we would have needed only 3 latent variables in the model. Instead the best model needs 6 LVs. Figure 4 shows a plot of actual vs. predicted concentration of polystyrene using the model based on μ_a. It can be seen that the accuracy of predictions drops with the increasing concentration of particles.

These observations could be explained by the insight provided by taking a closer look at the bulk absorption coefficient given by (4). Although μ_a is free from nonlinear photon pathlength variations, it still has some variation not related to chemical information. The information about the concentration of the scattering component is contained in the term $\mu_{ap} = \sigma_{ap} c_p$ in (4). It is the only problematic term in μ_a from the point of view of multivariate calibration, because it varies not only with volumetric concentration of particles but also with their morphology (size and shape).

To examine this term further, for spherical particles it can be rewritten as:

$$\boldsymbol{\mu_{ap}} = \boldsymbol{\sigma_{ap}} \cdot c_p = \frac{\boldsymbol{\sigma_{ap}}}{V_p} \cdot c_p^{\#} = \frac{3 \cdot \boldsymbol{\sigma_{ap}}}{4 \cdot \pi \cdot R^3} \cdot c_p^{\#} = \boldsymbol{K} \cdot c_p^{\#} \quad (9)$$

V_p is the volume of a single particle (cm^3), c_p is the concentration of particles expressed as number density (cm^{-3}), $c_p^{\#}$ is the volumetric concentration of particles (ml/ml) and R is the particle radius. The term separated from $c_p^{\#}$ is denoted as $\boldsymbol{K}$. The parameter $\boldsymbol{K}$ is a function of particle radius both explicitly as well as implicitly due to σ_{ap} also being a function of particle radius[17]. While it does not explicitly contain concentration information, due to its dependence on particle size, it will be correlated to the concentration of particles. This is because particle concentration can be changed in 3 different ways: by keeping the number density of the particles the same and changing their size (radius), keeping the particle size the same and changing the number density or by a combination of both. From (9), looking at the first right-hand-side relation, it can be seen that the effect due to particle number density is a baseline offset of the absorption cross-section of the particle, whereas the effect due to particle size is manifested by a wavelength dependent change in the particle absorption cross-section. When both vary simultaneously, the fact that the number density is implicitly related to the particle size and thus indirectly to the absorption cross-section and the multiplicative (confounding) effect on one another means that only a portion of the concentration information can be extracted from the combined effect. Thus variations arising from this term can have an adverse affect on model performance when the models are built for predicting the concentration of a particulate species (i.e. a species that both absorbs and scatters light). The fact that we need three extra latent variables to predict the particle concentration may be due to requiring extra LVs to describe the effects described above.

Further, the accuracy of multivariate calibration models in predicting the concentration of the scattering component will depend on the magnitude of variation in $\boldsymbol{K}$. Since for the same particle concentration, $\boldsymbol{K}$ can take on different values due to changes in particle size, it introduces an error due to the confounding effect arising from the multiplicative nature of this parameter with respect to the particle concentration. The higher the concentration of particles, the larger will be the effect on $\boldsymbol{\mu_{ap}}$ due

to variations in $\boldsymbol{K}$. Since variations in $\boldsymbol{K}$ degrade the performance of a calibration model due to the confounding effect it induces, this translates into higher levels of uncertainty in the concentration estimates with increasing concentrations. This would explain the larger spread in the data in Figure 4 at higher concentrations.

EXPERIMENT

Simulation results show that a significant improvement in prediction accuracy can be achieved if PLS models were built on the bulk absorption coefficient rather than directly on reflectance or transmittance measurements, which are subject to nonlinear variations due to different pathlengths travelled by photons. To validate the concept, the proposed approach was applied to a simple turbid system comprising polystyrene particles suspended in water. The three measurements (diffuse reflectance, diffuse transmittance and collimated transmittance) and the extracted bulk absorption coefficient are presented in figure 5. As was seen with simulations, the magnitudes of variation are much larger in the measurements (diffuse reflectance, diffuse transmittance and collimated transmittance) than in the extracted $\mu_{\mathbf{a}}$. Again, the majority of undesirable variation occurring due to physical effects has been successfully removed by extracting the bulk absorption coefficient $\mu_{\mathbf{a}}$.

To check the consistency in extracted μ_a, its profiles for all seven concentrations for a single particle size (430nm diameter) were plotted (figure 6). The differences in μ_a of the seven samples were only due to changes in polystyrene concentration. If the extraction step was effectively carried out, the peak where polystyrene absorbs should systematically increase with the concentration of polystyrene. While this was true with the 100nm particles, for samples with particle sizes larger than 100nm the bulk absorption spectra for the lowest two concentrations of polystyrene did not fall in the right order. In figure 6, this can be seen for the samples with polystyrene particles of 430nm diameter. This is likely due to the increased losses of light in the integrating sphere setup. The adding-doubling method does not take into account the light lost through the sides of the cuvette. When the particle number density is low, the mean free path of the photons becomes high. As a result a photon that is scattered sideways has

a greater probability of reaching the side walls of the cuvette since the probability of it getting scattered again before it reaches the wall becomes lower. Since the adding-doubling method assumes that the breadth of the cuvette is infinite, any loss through the side walls is manifested as absorption. When this effect becomes significant, the bulk absorption coefficient extracted using the adding-doubling method is overestimated. The reason it is evident for the larger particles is because for the same volumetric concentration of particles, there are much fewer number of large particles. This is due to the fact the number of particles is related to the cube of the particle radius when the total particle volume is kept constant. This explains why the absorption (μ_a) spectra for the lowest concentrations were shifted up and had higher values. To correct these offsets in μ_a due to light losses, EMSCL was applied. It was found that the application of EMSCL successfully corrected the baseline variations introduced by light loss from the sides of the cuvette. This is evident in figure 7 where it is seen that in the region of polystyrene absorption, the EMSCL corrected absorption spectra now shows an increase in μ_a with increasing concentration of polystyrene latex particles.

Having obtained consistent estimates of bulk absorption coefficient after pre-processing with EMSCL, a PLS calibration model was built on the pre-processed bulk absorption spectra. For all three cases the EMSCL technique provided the best (or the same level of) performance and therefore only this case is reported for the models built on μ_a, R_d and T_d. The RMSECV curves of calibration models which gave the lowest prediction errors are given in figure 8(a) and the predicted vs. actual values are shown in figures 8(b)-(d). The results are summarized in table 3.

The system considered here consisted of two components. Therefore, theoretically, one latent variable should have been sufficient to model it. However, it is apparent from the RMSECV curve of μ_a that three latent variables are needed to describe variation in it. This result agrees with the conclusion drawn from the analysis of simulated data that extra LVs are needed to describe the nonlinear variation in absorption coefficient of particles. The other important finding in the analysis of simulated data that the

prediction accuracy drops with increasing particle concentration is not so clear from the predicted versus actual values plot of experimental data (figure 8(b)).

It is also seen that while the proposed approach considerably outperformed the model obtained from diffuse reflectance spectra with EMSCL applied to it, very surprisingly the models using diffuse transmittance performed much better than the proposed approach. From physical considerations and the results from simulated data with 4 components this appears to be contradictory. This apparently contradictory result can be explained by examining the PLS scores and loadings plots for models using $\boldsymbol{T_d}$, $\boldsymbol{R_d}$ and $\boldsymbol{\mu_a}$.

Figures 9(a), (c) and (e) show the PLS loadings for the first three latent variables for models built using $\boldsymbol{\mu_a}$, $\boldsymbol{R_d}$ and $\boldsymbol{T_d}$, and figures 9(b), (d) and (f) show the corresponding scores of the first latent variable (LV1) plotted versus the particle concentrations. For all three cases, it is seen that the first LV is essentially modelling the water absorbance. While the wavelength region considered (1600-1848nm) contains the first overtones of O-H stretching, bending and libration vibrations of water at around 1790nm and the peak due to O-H stretching and bending vibrations at around1900nm[20] (of which only the tail part below 1848nm is included here), the region below the 1790nm peak has non-zero absorbance even though it is relatively a flat and featureless baseline. This "baseline" absorption will increase or decrease with changes in the concentration of water. It is this part of the spectrum that appears to be modelled by this LV. The second LV contains the first overtone peak due to the C-H stretching vibrations of polystyrene which occurs around 1680nm. For the models using $\boldsymbol{R_d}$ and $\boldsymbol{T_d}$, there is a significant correlation between the scores of the first LV and the particle concentrations. The correlation is the strongest for $\boldsymbol{T_d}$. For $\boldsymbol{\mu_a}$ this correlation is very weak and it could be argued that it is almost insignificant. Thus, models using $\boldsymbol{R_d}$ and $\boldsymbol{T_d}$ benefit from the correlation of particle concentration with the water absorption in the first LV whereas this secondary correlation which is due to the use of a two-component system is not available for the model using $\boldsymbol{\mu_a}$. In a two component system, the concentrations of the two components are inversely correlated due to the closure condition. Therefore, if the first LV was representing water absorption, then the scores of the first LV should have shown a

negative correlation with the water absorption because increasing the concentration of particles would result in the decrease of water concentration (due to volume displacement). However, it is seen that the correlation is positive. This positive correlation can be explained if we consider pathlength variations occurring due to changes in the particle concentrations. When the particle concentration increases, multiple scattering increases which in turn increases the pathlength travelled by the photons. This increase in pathlength means that the photons will travel longer distances through the medium (which is predominantly water) resulting in the absorption due to water increasing with increasing particle concentrations and thereby leading to a positive correlation of particle concentration with water absorption which is represented by the first LV. The fact that the volume displacement effect which would have manifested as a negative correlation is not evident indicates that this effect is much smaller than the effect due to pathlength variation which generates a positive correlation. It appears that this effect is largest in the diffuse transmittance measurements and to a lesser degree in diffuse reflectance measurements for the range of concentrations considered in this study. Naturally, in the case where the extracted bulk absorption spectra μ_a are used for building the PLS model, since the pathlength effect is removed by applying the RTE, this secondary correlation is mostly eliminated. As a result, the scores of LV1 show almost no correlation with the particle concentrations.

From this discussion it can be concluded that the model built using μ_a is almost fully based on the actual polystyrene signal whereas those built on the direct measurements have a significant contribution from the pathlength effect which only for a two-component system gives rise to additional correlation with the particle concentrations and which in turn leads to an apparent advantage. It should be noted that even with this advantage the model based on R_d does not outperform the proposed approach probably due to the fact that the pathlength correlation is not as strong as is the case with T_d. The analysis presented shows that the extraction algorithm using the RTE to obtain the bulk absorption spectra is successful in effectively removing pathlength variations and providing essentially a pathlength normalised absorption spectra. The discussion presented here also suggests that the models based on the direct measurements will lead to much larger errors when applied to a multi-component

system where the secondary correlation will not exist. On the other hand, a model based on μ_a can be expected to have lesser deterioration when applied to multiple component systems provided the bulk absorption spectra could be extracted with similar levels of accuracy as in the present study.

In the current study, the methodology was applied to develop models for estimating the concentration of a particulate species that both absorbs and scatters light. It is common to find situations where the species of interest is purely absorbing and is dissolved in a matrix containing a mixture of absorbing and scattering components e.g. glucose in blood. Another case of interest is when the species of interest (purely absorbing) is adsorbed on the surface of scattering particles. This methodology is applicable to such cases too. From the point of removing multiple scattering effects through the extraction of μ_a, the procedure is unaffected by which of the three cases is being considered. The difference in each of the cases is how the species of interest contributes to μ_a and thus affect the calibration model built using this extracted property.

In theory, situations where the species of interest is purely absorbing and dissolved in a medium containing scatterers represent a comparatively simpler problem. Examining (4), if for example the purely absorbing species of interest is component 1, then the term $\sigma_{a,1} \cdot c_1$ varies only with concentration of species 1 as there is no particle size contribution to the absorptivity $\sigma_{a,1}$. The effect of particle size only occurs indirectly through the term representing the particulate species which is an additive term. When the (purely absorbing) species of interest is adsorbed on the particle, the situation could be expected to be slightly more complicated since the adsorbed species will modify the value of absorption cross-section σ_{ap} of the particle, the extent of which will depend on level of adsorption of the species. As a result, the number of latent variables required may be more than that indicated by the additive relationship given by (4).

CONCLUSIONS

The theoretical study using a simulated dataset containing 4 components with calibration models built for the case where the species of interest is a particle indicated that appreciable improvements in model

performance can be obtained when the proposed methodology is used compared to applying empirical scatter correction techniques to single measurements. For real systems, the ability to extract $\boldsymbol{\mu_a}$ consistently and with sufficient accuracy is key to fully realising the potential of this methodology. This in turn requires an accurate method for solving the RTE. In the current study, the adding-doubling (AD) method was used for this purpose. The methodology with the adding-doubling method as the engine for solving the RTE was tested with experimental data for a simple two-component (polystyrene-water) system. It was found system used in this study, the AD method introduced systematic errors in $\boldsymbol{\mu_a}$ due to light losses resulting from the finite width of the sample. To remove those errors additional pre-processing was required. Despite this drawback of the AD method, analysis indicates that the pathlength variations are effectively removed. It may be possible to obtain further improvements to this approach by minimising light loses by adjusting sample width and thickness.

The application of the methodology described here led to a model that performed better than that built using diffuse reflectance measurements. However its performance was worse compared to when diffuse transmittance measurements were used to build a model. Analysis indicates that this is due to the secondary correlation between particle concentration and pathlength travelled by the photons which occurs due to the system containing only two components. This secondary correlation will not be available in multi-component systems and therefore it can be argued that the performance of models built using $\boldsymbol{T_d}$ or $\boldsymbol{R_d}$ would show significant degradation compared to using $\boldsymbol{\mu_a}$.

ACKNOWLEDGMENT

This work was funded through Marie Curie FP6 (INTROSPECT) and EPSRC grants GR/S50441/01 and GR/S50458/01.

REFERENCES

[1] Reis, M. M.; Araujo, P. H. H.; Sayer, C.; Giudici, R. *Anal. Chim. Acta*. **2007**, *595*, 257-265.

[2] Gossen, P. D.; MacGregor, J. F.; Pelton, R. H. *Appl. Spectrosc.*, **1993**, *47*, 1852-1870.

[3] Vieira, R. A. M.; Sayer, C.; Lima, E. L.; Pinto, J. C. *J. Appl. Polym. Sci*. **2002**, *84*, 2670-2682.

[4] Reis, M. M.; Araujo, P. H. H.; Sayer, C.; Giudici, R. *Ind. Eng. Chem. Res*. **2004**. *43*, 7243-7250.

[5] Santos, A. F.; Silva, F. M.; Lenzi, M. K.; Pinto, J. C. *Polym. Plast. Technol. Eng*. **2005**, *44*, 1-61.

[6] Zeaiter, M.; Roger, J. M.; Bellon-Maurel, V. *J. Chemom. Intell. Lab. Syst*. **2006**, *80*, 227-235.

[7] Blanco, M.; Peinado, A. C.; Mas, J. *Anal. Chim. Acta*. **2006**, *556*, 364-373.

[8] Hjalmarsson, P.; Thennadil, S. N. In *Complex Dynamics and Fluctuations in Biomedical Photonics V*, Jan. 19-21, **2008,** San Jose CA. 6855, 85508-85508.

[9] Thennadil, S. N.; Martin, E. B. *J. Chemometr.*, **2005**, *19*, 77-89.

[10] Thennadil, S. N.; Martens, H.; Kohler, A. *Appl. Spectrosc*. **2006**, *60*, 315-321.

[11] Abrahamsson, C.; Johansson, J.; Andersson-Engels, S.; Svanberg, S.; Folestad, S. *Anal. Chem.* **2005**, *77*, 1055-1059.

[12] Ishimaru, A. *Wave Propagation and Scattering in Random Media*; IEEE Press, Oxford University Press: Oxford, 1997.

[13] Prahl, S. A. In *Optical Thermal Response of Laser Irradiated Tissue*; Welch, A. J.; Gemert, M. J. C. Eds.: Plenum Press, New York, 1995; pp 101–129.

[14] Wang, L. H.; Jacques, S. L.; Zheng, L. Q. *Comput. Meth. Prog. Bio*. **1995**, *47*(2), 13

[15] Saeys, W.; Velazco-Roa, M. A.; Thennadil, S. N.; Ramon, H.; Nicolai, B. M. *Appl. Opt.* **2008**, *47*, 908-919.

[16] Dzhongova, E.; Harwood, C. R.; Thennadil, S. N. *Appl. Spectrosc*. **2009**, *63*, 25-32.

[17] Velazco-Roa M. A.; Thennadil, S. N. *Appl. Opt*. **2007**, *46*, 3730-3735.

[18] Bohren, C. F.; Huffman, D. R. *Absorption and Scattering of Light by Small Particles*; Wiley-VCH: Germany, 2004.

[19] Segelstein, D. J. *The complex refractive index of water*; MS Thesis, University of Missouri-Kansas City, 1981.

[20] Kradgel, C.; Lee, K. A. In *Handbook of Near Infrared Analysis*; Burns, D. A.; Ciurczak E. W. Eds.: CRC Press, Boca Raton, 2008; pp 529-568.

LIST OF FIGURES

LIST OF TABLES

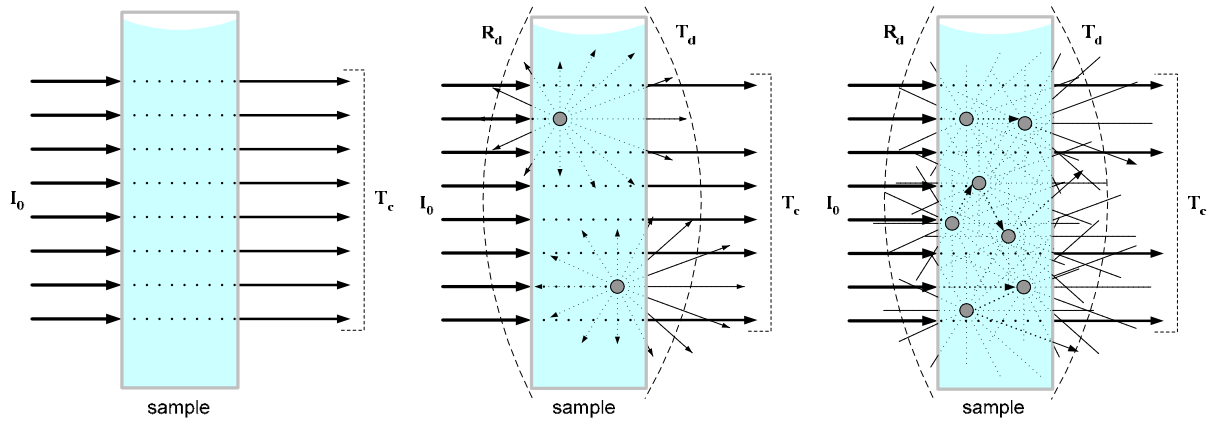

Figure 1. a.) Homogeneous sample (e.g. liquid mixture), absorption only; b.) Turbid sample with very low concentration of scatterers (particles) – absorption + single scattering; c.) Turbid sample with a high concentration of scatterers - absorption + multiple scattering.

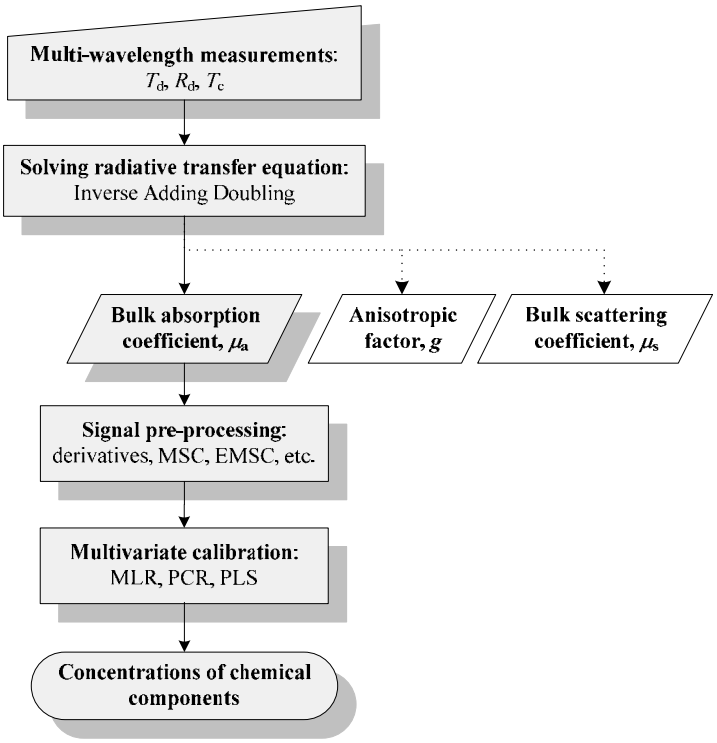

Figure 2. Flowchart illustrating the methodology used in this study for correcting multiple scattering effects and building calibration models.

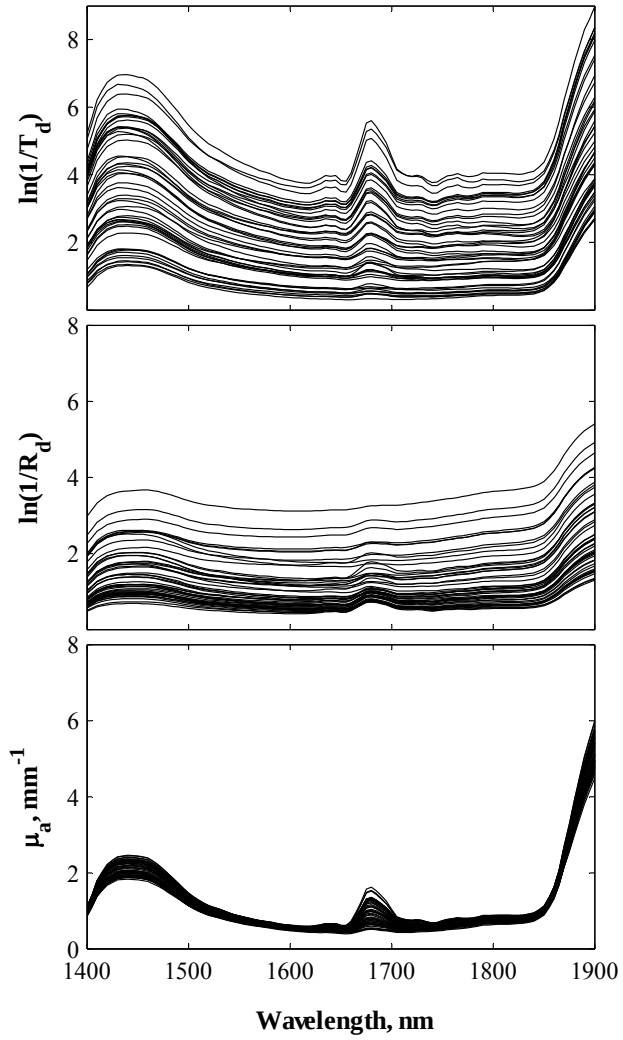

Figure 3. Total diffuse transmittance, total diffuse reflectance, and bulk absorption coefficient μ_a for calibration data set from simulations.

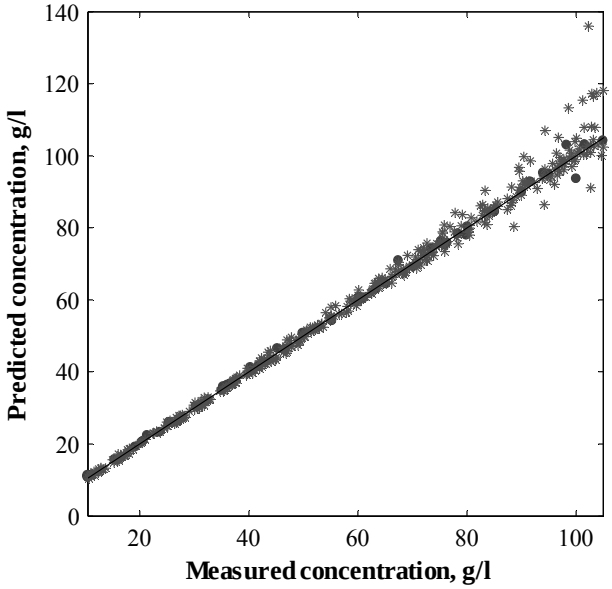

Figure 4. Predicted versus actual values of concentration of scattering component (polystyrene) for training and validation data sets (Simulations).

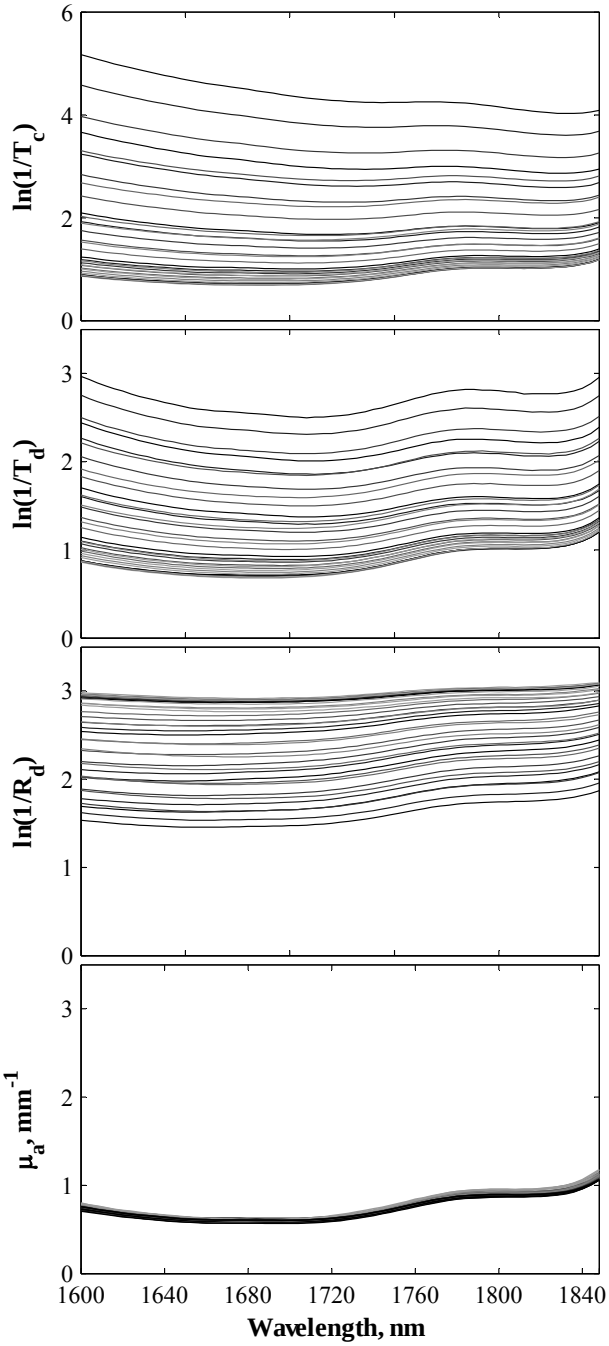

Figure 5. Experimental polystyrene-water data set: Collimated transmittance, total diffuse transmittance, total diffuse reflectance and bulk absorption coefficient μ_a.

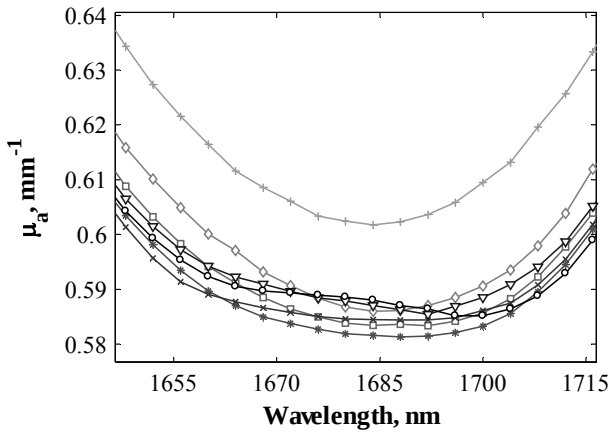

Figure 6. μ_a profiles of different concentrations for 430nm diameter polystyrene particles: 0.1%wt., 0.5%wt., 0.9%wt., 1.23%wt., 1.6%wt., 1.95%wt., 2.3%wt.

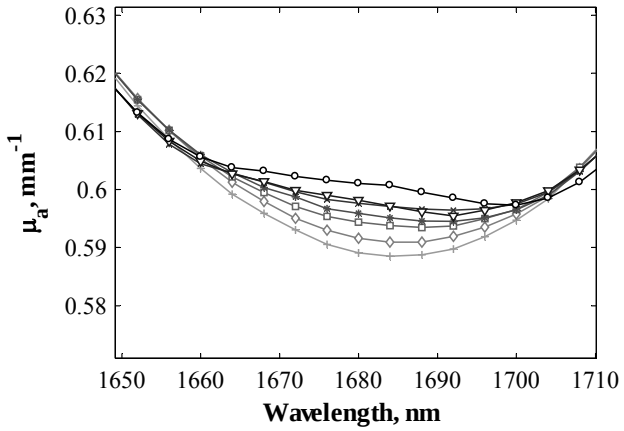

Figure 7. μ_a profiles of different concentrations after EMSCL: 0.1%wt., 0.5%wt., 0.9%wt., 1.23%wt., 1.6%wt., 1.95%wt., 2.3%wt.

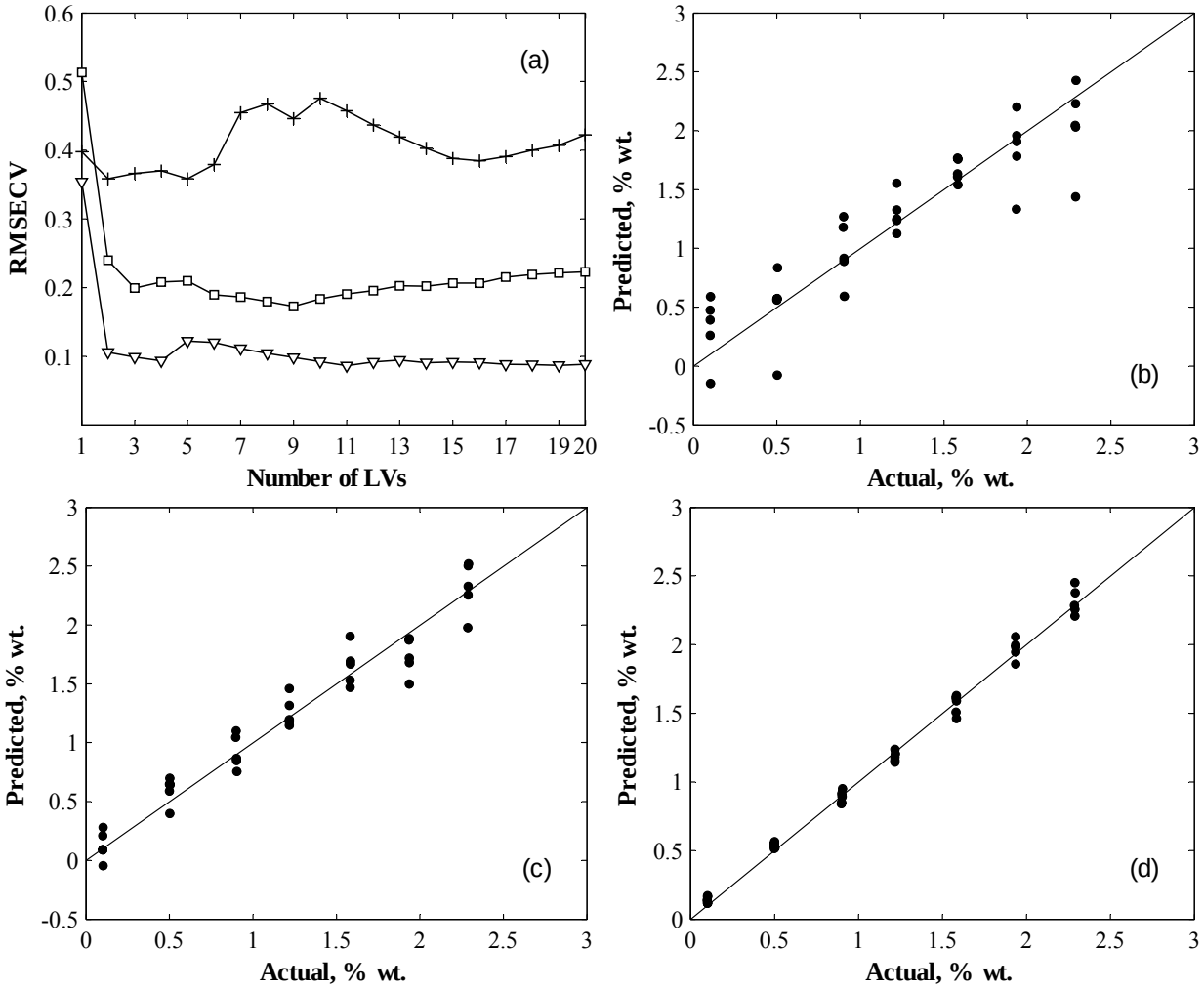

Figure 8. (a) RMSECV curves of different PLS calibration models built using: $\boldsymbol{R}_{\mathbf{d}}$ pre-processed with EMSCL, $\boldsymbol{T}_{\mathbf{d}}$ pre-processed with EMSCL, $\boldsymbol{\mu}_{\mathbf{a}}$ using EMSCL. Predicted concentration of polystyrene versus actual using: (b) $\boldsymbol{R}_{\boldsymbol{d}}$ +EMSCL, (c) $\boldsymbol{\mu}_{\mathbf{a}}$ +EMSCL and (d) $\boldsymbol{T}_{\boldsymbol{d}}$ + EMSCL.

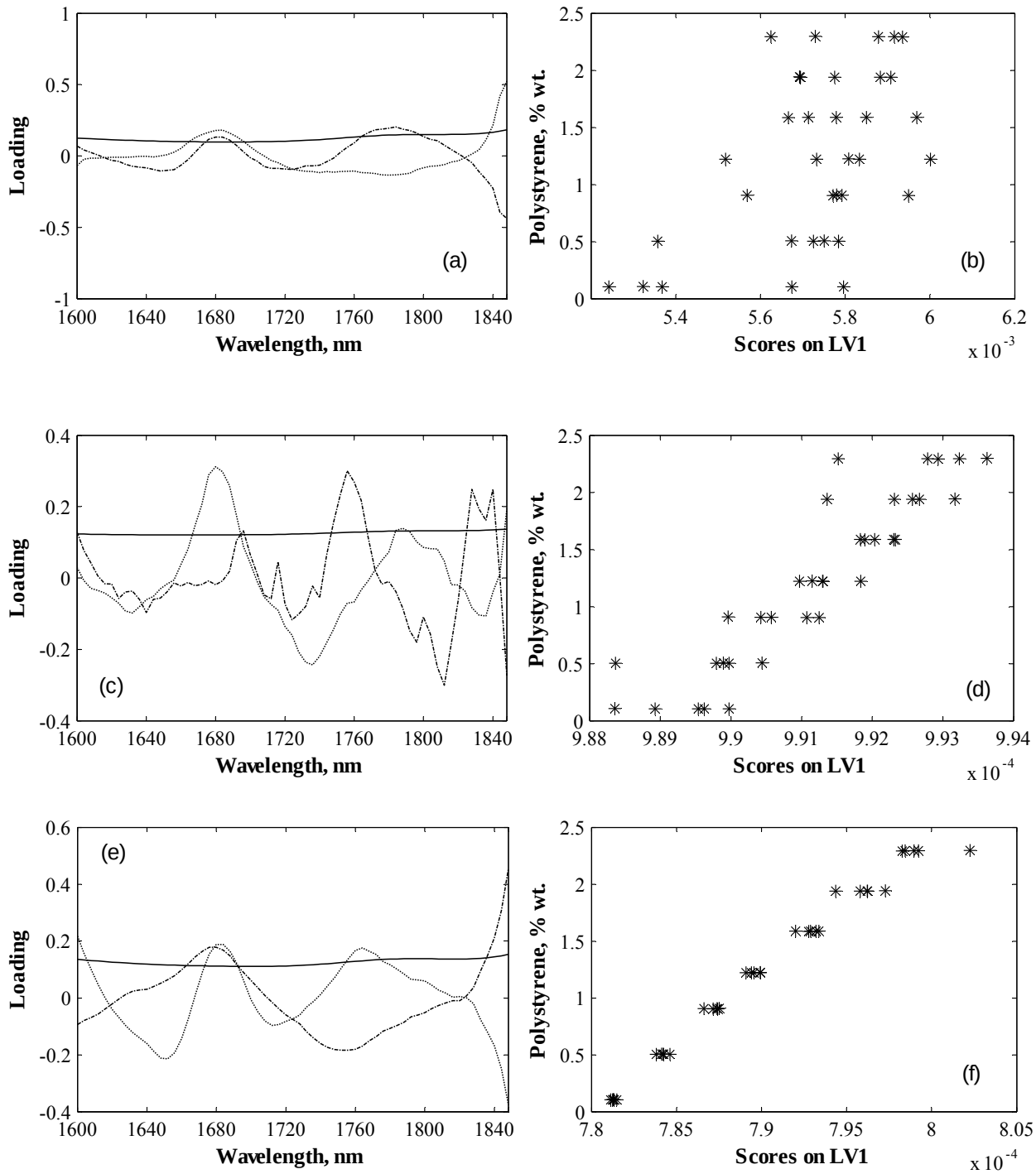

Figure 9. (a), (c) and (e) – Loadings of the first 3 LVs from PLS models obtained using $\boldsymbol{\mu_a}$, $\boldsymbol{R_d}$ and $\boldsymbol{T_d}$ respectively. Solid line LV1, -- LV2 and — – LV3. (b), (d) and (e) – Scores of LV1 vs. Particle concentrations for the PLS models obtained using $\boldsymbol{\mu_a}$, $\boldsymbol{R_d}$ and $\boldsymbol{T_d}$ respectively.

Table 1. Performance of calibration models for estimating polystyrene concentration in the simulated data-set of a four-component system.

Predictions of concentration of scattering component (polystyrene)				
	Preprocessing	LVs	Calibration	Test
			RMSECV (g/l)	RMSEP (g/l)
Calibration models built on diffuse reflectance	EMSCL	12	2.17	2.5
Calibration models built on diffuse transmittance	EMSCL	8	4.04	3.14
Calibration model built on μ_a	None	6	1.38	1.42

Table 2. Performance of calibration models for estimating polystyrene concentration in the experimental data-set of a two-component (polystyrene-water) system.

Model	Preprocessing	LVs	RMSECV
Calibration models built on total diffuse reflectance			
1	EMSCL	2	0.37
Calibration models built on total diffuse transmittance			
2	EMSCL	4	0.09
Calibration model built on μ_a			
3	EMSCL	3	0.23