Quantitative Spectroscopic Analysis of Heterogeneous Mixtures: the Correction of Multiplicative Effects Caused by Variations in Physical Properties of Samples

Jing-Wen Jin[a], Zeng-Ping Chen[a], Li-Mei Li[a], Raimundas Steponavicius[b], Suresh N. Thennadil[c], Jing Yang[a] and Ru-Qin Yu*[a]*

a. State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China

b. School of Chemical Engineering and Advanced Materials, Newcastle University, Merz Court, Newcastle upon Tyne, NE1 7RU, United Kingdom

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

* Corresponding author

Tel.: (+86) 731 88821916; Fax: (+86) 731 88821916;

E-mail Address: zpchen2002@hotmail.com (Z.P. Chen), rqyu@hnu.cn (R.Q. Yu)

ABSTRACT: Spectral measurements of complex heterogeneous types of mixture samples are often affected by significant multiplicative effects resulting from light scattering, due to physical variations (e.g. particle size and shape, sample packing and sample surface, etc.) inherent within the individual samples. Therefore, the separation of the spectral contributions due to variations in chemical compositions from those caused by physical variations is crucial to accurate quantitative spectroscopic analysis of heterogeneous samples. In this work, an improved strategy has been proposed to estimate the multiplicative parameters accounting for multiplicative effects in each measured spectrum, and hence mitigate the detrimental influence of multiplicative effects on the quantitative spectroscopic analysis of heterogeneous samples. The basic assumption of the proposed method is that light scattering due to physical variations has the same effects on the spectral contributions of each of the spectroscopically active chemical component in the same sample mixture. Based on this underlying assumption, the proposed method realizes the efficient estimation of the multiplicative parameters by solving a simple quadratic programming problem. The performance of the proposed method has been tested on two publicly available benchmark data sets (i.e. near-infrared total diffuse transmittance spectra of four-component suspension samples and near infrared spectral data of meat samples) and compared with some empirical approaches designed for the same purpose. It was found that the proposed method provided appreciable improvement in quantitative spectroscopic analysis of heterogeneous mixture samples. The study indicates that accurate quantitative spectroscopic analysis of heterogeneous mixture samples can be achieved through the combination of spectroscopic techniques with smart modeling methodology.

Keywords: Heterogeneous mixture samples, Multiplicative light scattering effects, Modified optical path-length estimation and correction, Dual calibration strategy, Spectroscopic quantitative analysis

1. Introduction

The quantitative analysis of heterogeneous mixture samples using conventional instruments such as HPLC generally involves troublesome and time-consuming sample preparations. Due to their high measuring speed, multiplicity of analysis, non-destructivity, flexibility and especially requirement of less or even no sample preparations, spectroscopic technologies such as near infrared (NIR), mid infrared (MIR) and Fourier-transform Raman spectroscopy (FT-Raman) have been increasingly applied to the analysis of complex systems in areas of chemicals, food processing, agriculture and pharmaceuticals, etc [1-6]. However, when analyzing complex heterogeneous mixture samples that exhibit sample-to-sample variability in physical properties using spectroscopic instrumentation, the multiplicative light scattering effects caused by the uncontrolled variations in optical path length due to the physical differences between samples (e.g. particle size and shape, sample packing, and sample surface, etc) would 'scale' the entire spectral measurement and hence mask the spectral variations relating to the content differences of chemical compounds in the samples [7]. The presence of dominant multiplicative effects in spectral data could invalidate the underlying assumption of commonly used multivariate linear calibration methods such as PCR [8] and PLS [9] which postulates a linear relationship between spectral measurements and the contents of chemical components, and hence significantly deteriorate the predictive performance of calibration models built by multivariate linear calibration methods. The separation of the spectral contributions due to variations in chemical compositions from those caused by multiplicative effects is therefore crucial to the accurate quantitative analysis of messy spectral data with multiplicative effects.

A number of chemometric pre-processing methods, e.g., Multiplicative Signal Correction (MSC) [7], Standard Normal Variate (SNV) [10], Inverted Signal Correction (ISC) [11], Extended Inverted Signal Correction (EISC) [12], Extended MSC (EMSC) [13] and Modified EMSC [14] have been proposed to remove the multiplicative effects caused by variations in physical properties of samples. However MSC, ISC and EISC could only be applied to a spectrum that has wavelength regions containing no chemical information, i.e. influenced only by the multiplicative effects. Otherwise, they could result in dramatically poor results. The applicability of EMSC and the modified EMSC is limited due to the requirement of the pure spectra for all spectroscopically active chemical components present in the samples which is difficult to satisfy in practice.

Recently, Thennadil et al. proposed an interesting approach for the correction of multiple light scattering effects by making use of radiative transfer theory [15-16]. Though this approach can to some extent improve the predictive performance of multivariate calibration models, its implementation complexity and the requirement of three measurements for each mixture sample (i.e. total diffuse transmittance, total diffuse reflectance and collimated transmittance) make it difficult to use in practice. More recently in a review of pharmaceutical applications of separation of absorption and scattering in near-infrared spectroscopy, similar concepts to the approach mentioned above are discussed [17]. Another similar approach to compensate for the scattering effects in reflectance spectroscopy was developed by Kessler et al. by integrating Kubelka–Munk equation with multivariate curve resolution (MCR) [18]. Like the method based on radiative transfer theory, the application of hard model constrained MCR–ALS algorithm is dependent on the availability of two measurements for each mixture sample (i.e. the diffuse

reflectance spectra of a sample with an optically infinite thickness and a sample of finite thickness). Hence the scope of its applicability is also limited.

To overcome these limitations, one of the present authors developed a novel multiplicative effect correction approach, Optical Path-Length Estimation and Correction (OPLEC) [19, 20]. OPLEC adopted the following two-step procedure for the correction of multiplicative effects in spectral measurements. First of all, the multiplicative parameters accounting for multiplicative effects in the spectral measurements of the calibration samples are estimated by a unique method deduced solely from the linear transformation of the calibration spectral measurements. And then the multiplicative effects in the spectral measurements of the test samples are efficiently removed by a dual-calibration strategy. Without placing any requirement on the spectral measurements, OPLEC can efficiently separate the multiplicative effects of samples' physical properties from the spectral variations related to the chemical compositions, and hence has much wider applicability than other methods reported in the literature. The development of OPLEC provided an important contribution to the solution of multiplicative light scattering issues. Whereas the first step of OPLEC, i.e. the estimation of the multiplicative parameters for the calibration samples, involves the determination of the number of spectroscopically active chemical components in the systems under study. A poor estimation of the number of chemical components would result in suboptimal performance of OPLEC. For complex systems, the estimation of the number of chemical components is not a trivial task. Therefore, the OPLEC method needs to be refined to realize its full potential for spectroscopic quantitative analysis of heterogeneous mixtures.

The objectives of this study were (1) to redesign the method in OPLEC for the estimation of the multiplicative parameters for the spectral measurements of the calibration samples, (2) to develop a

104 simple but effective approach for determining the optimal model parameter (i.e. the number of
105 spectroscopically active chemical components) in OPLEC, (3) to improve the robustness of OPLEC
106 when being applied to complex systems, and finally (4) to evaluate the performance of the modified
107 OPLEC method on two publicly available benchmark data sets.
108
109
110 **2. Theory**
111 *2.1 The dual calibration strategy adopted by OPLEC to correct multiplicative effects*
112 For spectral measurements with multiplicative effects caused by changes in the optical path-length due
113 to the physical variations of the samples, the measured spectrum ($\mathbf{x}_i$, row vector) of sample i composed
114 of J chemical components can be approximated by the following model [6, 7, 21]:

$$\mathbf{x}_i = p_i \sum_{j=1}^{J} c_{i,j}\mathbf{s}_j, \quad i = 1,2,\ldots,I \qquad (1)$$

115 Where $c_{i,j}$ is the concentration of the j-th chemical component in the i-th mixture sample; $\mathbf{s}_j$ represents
116 the pure spectrum of j-th chemical component in the mixtures. The coefficient p_i accounts for the
117 multiplicative effects in the spectral measurements of the i-th sample caused by changes in the optical
118 path-length due to the physical variations of the sample; I denotes the number of calibration samples.
119 Assume the first component is the target constituent in the mixtures and $\sum_{j=1}^{J} c_{i,j}\mathbf{s}_j = 1$ (which strictly
120 hold for $c_{i,j}$ representing unit-free concentration such as weight fraction and mole fraction), then eq.1
121 can also be expressed as:

$$\mathbf{x}_i = p_i c_{i,1} \Delta \mathbf{s}_1 + p_i \mathbf{s}_2 + \sum_{j=3}^{J} p_i c_{i,j} \Delta \mathbf{s}_j, \quad \Delta \mathbf{s}_j = \mathbf{s}_j - \mathbf{s}_2 \tag{2}$$

It is obvious that a linear relationship exists between $\mathbf{x}_i$ and p_i, and also between $\mathbf{x}_i$ and $p_i c_{i,1}$. It should be noted that this conclusion would also hold when the content of one constituent (or matrix substances) does not vary over mixture samples. Provided the multiplicative parameter vector $\mathbf{p}$ ($\mathbf{p} = [p_1; p_2; \ldots; p_I]$) for the calibration samples is available (actually it can be estimated from the calibration spectra by the multiplicative parameter estimation method outlined in section 2.2) , two following calibration models can therefore be built by multivariate linear calibration methods such as PLS. The first model is between $\mathbf{X}$ ($\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; \ldots; \mathbf{x}_I]$) and $\mathbf{p}$, and the other is between $\mathbf{X}$ and diag($\mathbf{c}_1$)$\mathbf{p}$ ($diag(\mathbf{c}_1)\mathbf{p} = [p_1 \times c_{1,1}; p_2 \times c_{2,1}; \ldots; p_I \times c_{I,1}]$). For simplicity, the same number of latent components is generally used in the above two PLS calibration models. Once the spectrum of a test sample has been recorded, the content of the target constituent in the test sample can then be obtained by dividing the prediction of the second calibration model by the corresponding prediction of the first calibration model.

2.2 Multiplicative parameter estimation

Obviously, the estimation of the multiplicative parameter vector $\mathbf{p}$ for the calibration samples is the key to the correction of the multiplicative effects by the above dual calibration strategy. The performance of the multiplicative parameter estimation method in the original OPLEC method [19] relies on the accurate estimation of the number of spectroscopically active chemical components in the systems under study. Poor estimation of the number of chemical components could significantly affect the performance of

OPLEC. With a view to improve the robustness of OPLEC, the following refined method for the estimation of multiplicative parameter vector **p** for the calibration samples was proposed in this work.

Suppose the singular value decomposition of **X** ($\mathbf{X}=[\mathbf{x}_1;\mathbf{x}_2;\ldots;\mathbf{x}_I]$) can be expressed as follows:

$$\mathbf{X}=[\mathbf{U}_s,\mathbf{U}_n]\begin{bmatrix}\sum_s & 0\\ 0 & \sum_n\end{bmatrix}[\mathbf{V}_s,\mathbf{V}_n]^{\mathrm{T}}=\mathbf{U}_s\sum_s\mathbf{V}_s^{\mathrm{T}}+\mathbf{E} \quad (3)$$

Where, $\mathbf{E}=\mathbf{U}_n\sum_n\mathbf{V}_n^{\mathrm{T}}$; superscript 'T' denotes the transpose; subscripts 's' and 'n' signify that the corresponding factors represent spectral information and noise, respectively. Suppose the actual number of spectroscopically active chemical components in the system studied is *r*, then both $\mathbf{U}_s$ and $\mathbf{V}_s$ consist of *r* columns. According to eq.2, both vectors **p** and diag($\mathbf{c}_1$)**p** are in the column space of $\mathbf{U}_s$, so the following equations hold:

$$\mathbf{U}_s\mathbf{U}_s^{\mathrm{T}}\mathbf{p}=\mathbf{p} \quad (4)$$

$$\mathbf{U}_s\mathbf{U}_s^{\mathrm{T}}diag(\mathbf{c}_1)\mathbf{p}=diag(\mathbf{c}_1)\mathbf{p} \quad (5)$$

Since there is no requirement to know the absolute value of p_i, p_i can be assumed to be no less than unity ($\mathbf{p}\geq\mathbf{1}$). Therefore, the vector **p** satisfying equations 4 and 5 can be obtained by solving the following constrained optimization problem:

$$\min_{\mathbf{p}}\frac{1}{2}\left(\left\|\mathbf{U}_s\mathbf{U}_s^{\mathrm{T}}\mathbf{p}-\mathbf{p}\right\|_2^2+\frac{1}{w^2}\left\|\mathbf{U}_s\mathbf{U}_s^{\mathrm{T}}diag(\mathbf{c}_1)\mathbf{p}-diag(\mathbf{c}_1)\mathbf{p}\right\|_2^2\right),\ \text{subject to}\ \mathbf{p}\geq 1 \quad (6)$$

Where, $\|\ \|_2$ denotes l^2 norm; w is a weight to balance the two parts in the above optimization function. It can be simply set to be the maximum element of $\mathbf{c}_1$. The above constrained optimization problem can

be transformed into an equivalent quadratic programming problem (which can be resolved by the *quadprog* function in MATLAB. The MATLAB code for the multiplicative parameter estimation method is available in Supporting Information):

$$\min_{\mathbf{p}} f(\mathbf{p}) = \frac{1}{2}\mathbf{p}^{\mathrm{T}}((\mathbf{I}-\mathbf{U}_s\mathbf{U}_s^{\mathrm{T}}) + diag(\mathbf{c}_1/w)(\mathbf{I}-\mathbf{U}_s\mathbf{U}_s^{\mathrm{T}})diag(\mathbf{c}_1/w))\mathbf{p}, \text{ such that } -\mathbf{p} \le -1 \quad (7)$$

2.3 Determination of the number of columns in $\boldsymbol{U}_s$

Theoretically, the number of columns in $\mathbf{U}_s$ (i.e. parameter *r*) should equal to the number of spectroscopically active chemical components in the systems under study. It is generally difficult to determine the exact number of spectroscopically active chemical components in a complex system. Moreover, when the spectral data does not strictly obey the model in eq. 1, the optimal number of columns in $\mathbf{U}_s$ might not solely depend on the number of spectroscopically active chemical components in the system under study, which would further complicate the situation. Fortunately, a simple mathematical analysis reveals that $\min_{\mathbf{p}} f(\mathbf{p})$ decreases dramatically with the increase of *r* at the very start, and then tends to be steady when *r* exceeds certain threshold value. Therefore, the optimal value of *r* can be determined by locating the turning point in the plot of $\min_{\mathbf{p}} f(\mathbf{p})$ versus *r*.

3. Case studies

The effectiveness of the modified OPLEC method (hereafter referred to OPLEC_m) with respect to its ability to estimate multiplicative parameters was first tested on the near-infrared total diffuse transmittance spectra of four-component suspension system consisting of water, deuterium, ethanol, and

polystyrene (hereafter referred to four-component suspension data). To further explore the potential of $OPLEC_m$, another real-world near-infrared transmittance spectra of meat samples recorded on a Tecator Infratec Food and Feed Analyzer (hereafter referred to tecator data) is employed. This spectral data set is publicly available and hence ensures that the interested reader can repeat the analysis.

3.1 Four-component suspension data[16]

The four-component suspension system is composed of three fully miscible absorbing species of water, deuterium oxide and ethanol and a species that both absorbs and scatters light (i.e., a particulate species of polystyrene). Specifically, the range of particle size and concentration were chosen to be 100~500 nm and 1~5 wt%, respectively, such that the following conditions were satisfied: stable suspension, multiple scattering, and sufficient signals in measurement. A total of 42 samples were prepared using various combinations of the concentrations of the four components and particle sizes of which the total diffuse transmittance ($\mathbf{T}_d$) spectra were recorded on a scanning spectrophotometer (CARY 5000) fitted with a diffuse reflectance accessory (DRA-2500). The spectral data were collected in the wavelength region of 1500-1880 nm with an interval of 2nm, resulting in measurements at 191 discrete wavelengths per spectrum. Twenty-two suspension samples' spectra were randomly selected to construct the calibration data set. The remaining twenty spectra from the other suspension samples made up the test data set. The absorbing-only species of deuterium oxide with concentration range between 20% and 58 wt% was taken as the analyte of interest in the present analysis and all the total diffuse transmittance spectra were transformed into absorbance spectra prior to the analysis. More experimental details can be found in the original paper of Steponavicius and Thennadil[16].

194

195 *3.2 Tecator data*[22]

196 This benchmark spectral data set consists of the near-infrared absorbance spectra of 240 meat samples
197 recorded on a Tecator Infratec Food and Feed Analyzer working in the wavelength range 850-1050 nm
198 with an interval of 2nm by the Near Infrared Transmission principle. Each sample contains finely
199 chopped pure meat with different moisture, fat and protein contents. A Soxhlet method was used as the
200 laboratory reference for fat determination. The Soxhlet values ranged from 2% to 59% fat. The 240
201 spectra were divided into 5 data sets for the purpose of model validation and extrapolation studies
202 (calibration set: 129; validation set: 43; test set: 43; extrapolation set for fat: 8; extrapolation set for
203 protein: 7). The task in the present work is restricted to predict the fat content (%) of a meat sample on
204 the basis of its near infrared absorbance spectrum, the extrapolation set for protein is therefore excluded.
205 The tecator data is available at http://lib.stat.cmu.edu/datasets/tecator.

206

207 *3.3 Data pre-treatment*

208 For the aforementioned two data sets, the possible additive baseline effects and wavelength dependent
209 spectral variations were firstly removed by projecting the measured spectra onto the orthogonal
210 complement of the space spanned by the row vectors of $\mathbf{M} = [\mathbf{1}; \boldsymbol{\lambda}; \boldsymbol{\lambda}^2]$ [19]. The pre-processed spectra
211 were then used to calculate the multiplicative parameter vector **p** for the calibration samples. The dual
212 calibration models in $OPLEC_m$ were built on the pre-processed spectra by using PLS method. The
213 predictive performance of $OPLEC_m$ was compared with those of PLS calibration models with and
214 without the application of data preprocessing methods such as MSC, SNV, EISC and EMSC as long as

215 they are applicable. The root-mean-square error of prediction (RMSEP) was used to assess the
216 performance of the calibration models.
217
218
219 **4. Results and discussion**
220 *4.1 Four-component suspension data*
221 The raw total transmittance spectra of the four-component suspension samples are presented in Figure 1.
222 It can be observed that the variations in polystyrene particle size and concentration across samples
223 resulted in significant additive baseline shift as well as multiplicative effects in the spectral data.
224 Though the additive baseline effects and possible wavelength dependent spectral variations can be
225 readily removed by orthogonal projection pre-processing, the multiplicative effects as a consequence of
226 the changes in sample's effective optical path-length are rather difficult to correct. Such multiplicative
227 effects can not be effectively modeled by multivariate linear calibration models either. Without being
228 properly corrected or modeled, they can significantly deteriorate the predictive performance of
229 multivariate linear calibration models [13, 19].

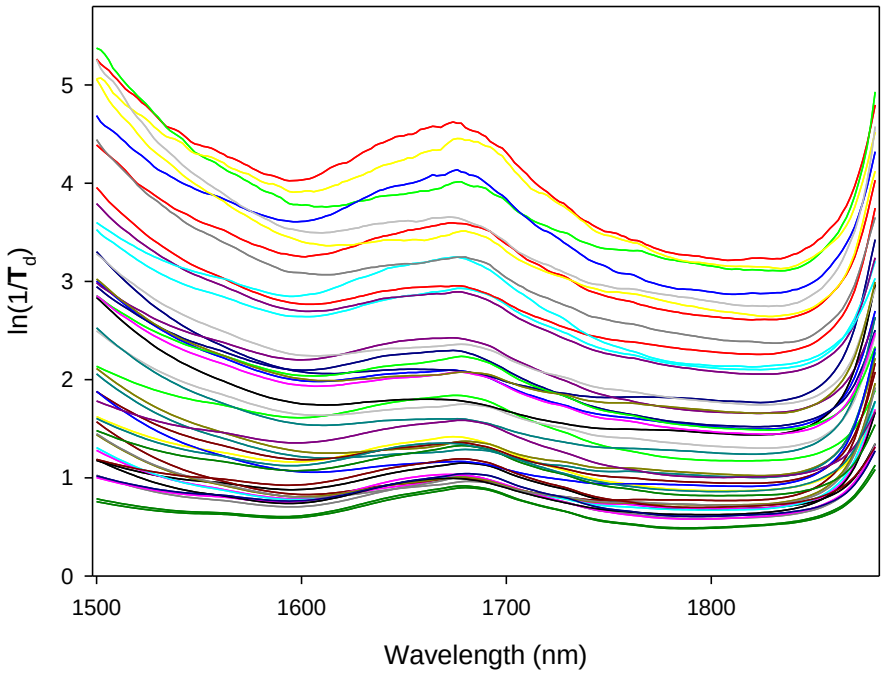
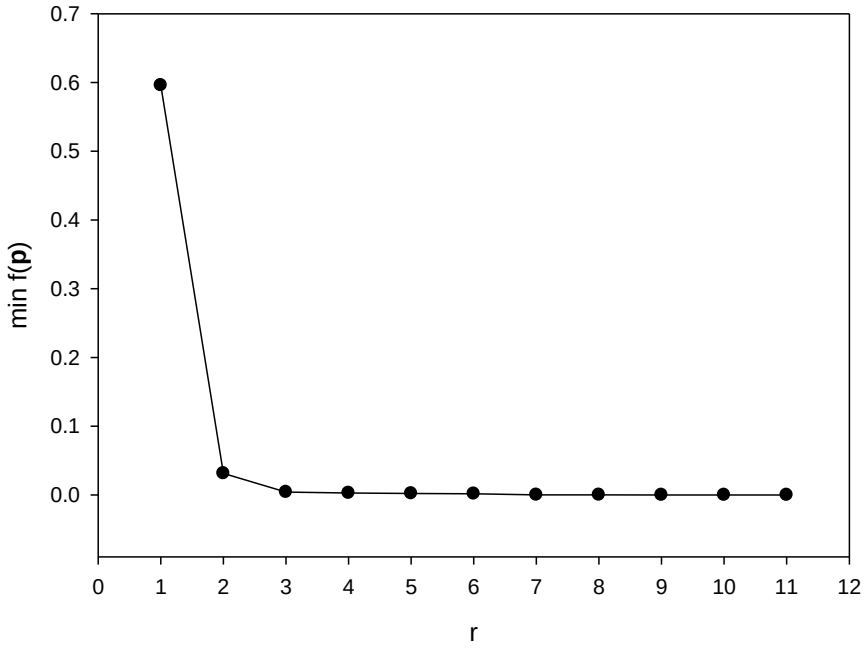


Figure 1: The raw spectra of the four component suspension system.

As stated in the theory section, $OPLEC_m$ can effectively correct the multiplicative effects in spectral measurements. $OPLEC_m$ consists of two main steps. The first step is to estimate the multiplicative parameter vector $\mathbf{p}$ for the calibration samples from the orthogonal projection pre-processed spectra. The estimation of the multiplicative parameter vector $\mathbf{p}$ for the calibration samples requires the determination of the actual number of spectral variation sources (r) in the calibration spectra, which can be achieved by scrutinizing the plot of $\min_{\mathbf{p}} f(\mathbf{p})$ versus r (Figure 2). From Figure 2, it can be seen that $\min_{\mathbf{p}} f(\mathbf{p})$ decreases obviously when the number of columns of $\mathbf{U}_s$ increases from one to three and including more components in $\mathbf{U}_s$ leads to no significant changes in $\min_{\mathbf{p}} f(\mathbf{p})$, which means the most spectral information relevant to $\mathbf{p}$ and $\mathrm{diag}(\mathbf{c}_1)\mathbf{p}$ was included in the first three principal components of $\mathbf{U}_s$. Therefore, the optimal value of r was then set to three.



244

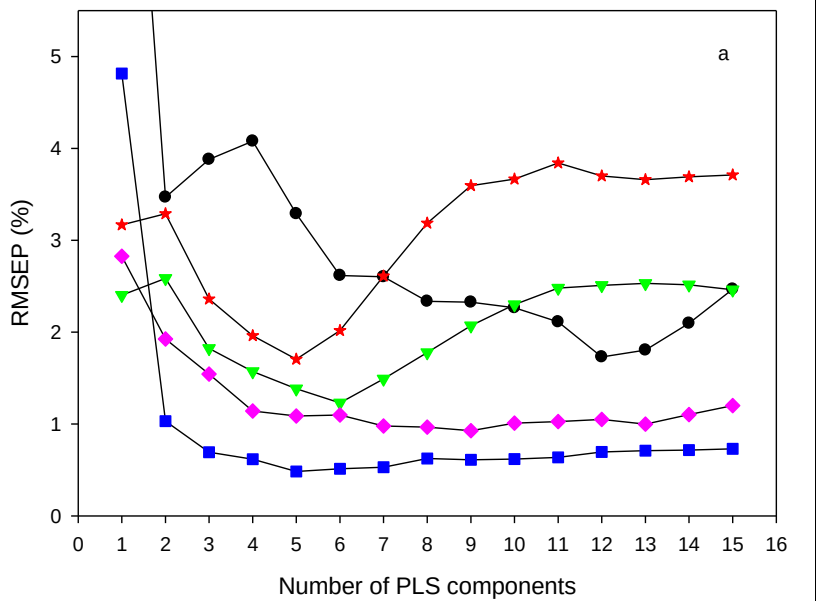
245 Figure 2: The relationship between $\min_{\mathbf{p}} f(\mathbf{p})$ and the number of columns of $\mathbf{U}_s$ (i.e. r) for the four

246 component suspension data.

After the estimation of the multiplicative parameter vector **p** for the calibration samples, one can assess the applicability of OPLEC$_m$ to the spectral data set by examining the two plots of **p** vs $\mathbf{U}_s\mathbf{U}_s^T\mathbf{p}$ and $diag(\mathbf{c}_1)\mathbf{p}$ vs $\mathbf{U}_s\mathbf{U}_s^T diag(\mathbf{c}_1)\mathbf{p}$, respectively (supporting information, Figure S-1). As shown in Figure S-1, both **p** and $diag(\mathbf{c}_1)\mathbf{p}$ are in good agreement with $\mathbf{U}_s\mathbf{U}_s^T\mathbf{p}$ and $\mathbf{U}_s\mathbf{U}_s^T diag(\mathbf{c}_1)\mathbf{p}$, respectively, which confirms that a linear relationship exists between $\mathbf{x}_i$ and p_i, and also between $\mathbf{x}_i$ and $p_i c_{i,1}$. The dual calibration strategy of OPLEC$_m$ is therefore applicable to the four component suspension data. Figure S-1 also reveals the presence of significant variations of multiplicative effects (p_i varying from 1 to 3.09) in the calibration samples. Multiplicative effect correction methods such as OPLEC$_m$ are therefore needed to remove such significant multiplicative effects in the spectral measurements.

Figure 3a compared the predictive performance of the optimal OPLEC$_m$ calibration model for deuterium oxide and the corresponding optimal PLS models with and without the application of preprocessing methods (e.g. SNV, MSC, EISC and EMSC). Obviously, as a result of the presence of severe multiplicative effects, PLS calibration model built on the raw calibration spectra could not give satisfactory predictions for the deuterium oxide in the test suspension samples. Preprocessing the calibration spectra by MSC, SNV or EISC can, to some extent, improve the predictive performance of PLS calibration models in terms of RMSEP values. However, due to the lack of a wavelength region containing no chemical information in the spectral data, the multiplicative effects can not be fully corrected by MSC, SNV or EISC. Hence, the predictive errors of the PLS calibration models built on the calibration spectra pre-processed by MSC, SNV and EISC are still comparatively high. As expected, OPLEC$_m$ offers the best improvement in terms of the predictive ability among all the pre-processed

methods. The $OPLEC_m$ calibration model with five underlying components provided the best predictive results with a $RMSEP_{test}$ value as low as 0.005, while the corresponding best $RMSEP_{test}$ value of the PLS calibration model with nine underlying components on the calibration spectra pre-processed by EISC is 0.009. Furthermore, the performance of the $OPLEC_m$ is robust to the number of columns in $\mathbf{U}_s$ (Figure 3b). Considering the fact that $OPLEC_m$ does not place any extra requirement on the spectral measurements as other multiplicative effect correction methods do, such a result is quite encouraging.



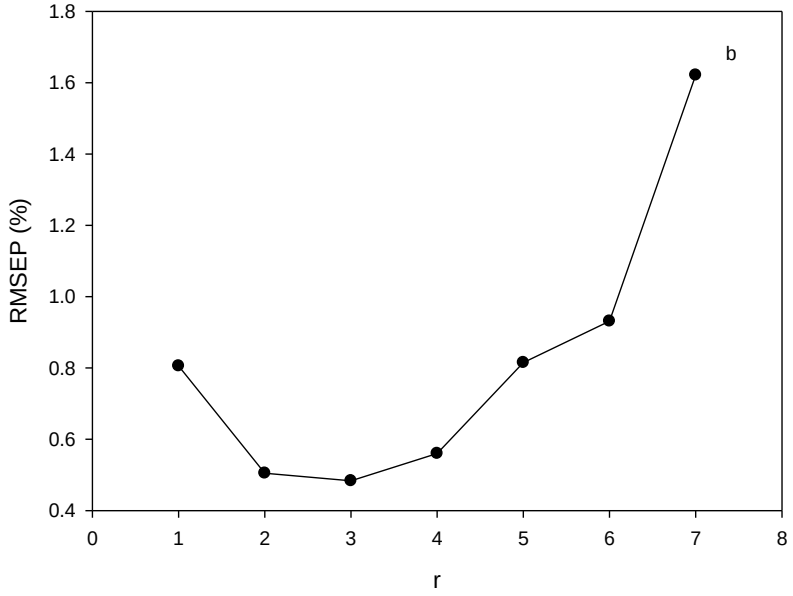
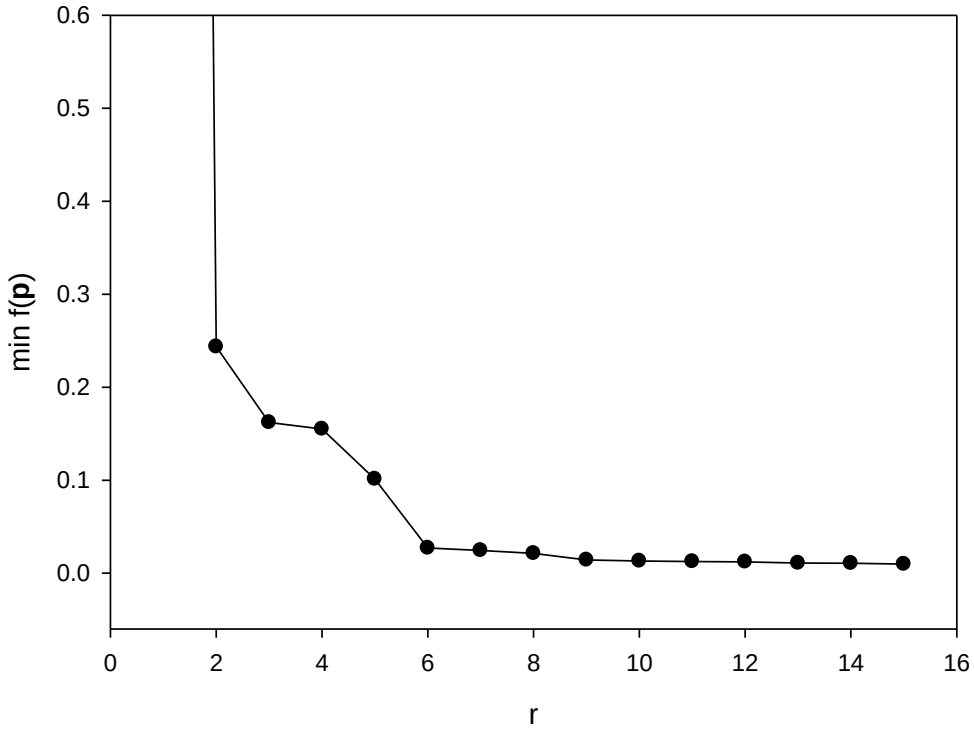


Figure 3: a) The predictive performance of $OPLEC_m$ and the PLS models built on the calibration spectra of the four component suspension system preprocessed by different methods (black circle: the raw spectra; red star: MSC; green triangle down: SNV; pink diamond: EISC; blue square: $OPLEC_m$); b) The predictive performance of the optimal $OPLEC_m$ models when $\mathbf{U}_s$ with different number of columns (r) were used in the calculation of the multiplicative parameter vector $\mathbf{p}$ for the calibration spectra.

4.2 Tecator data

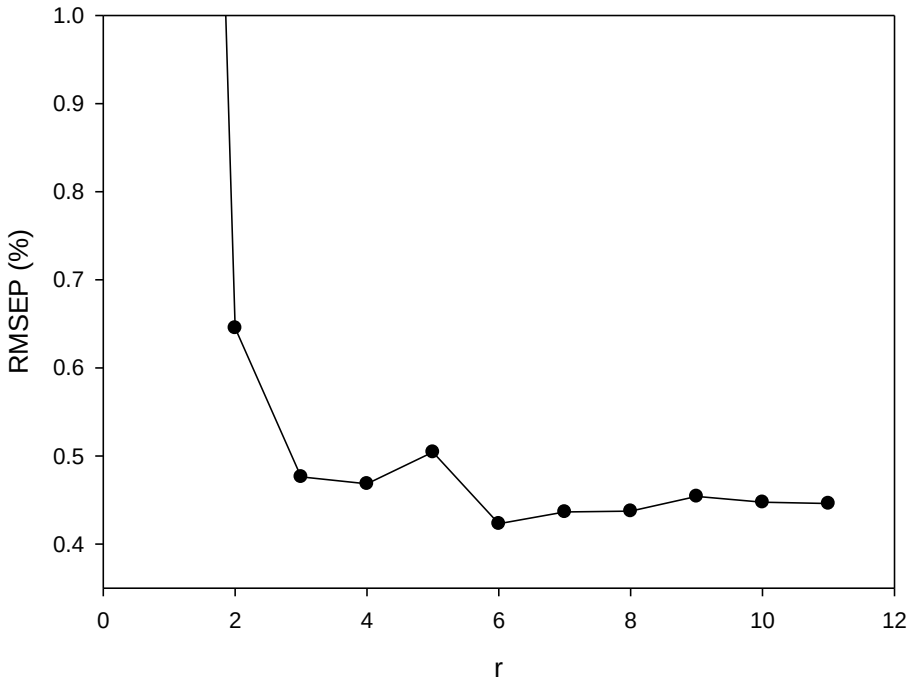
As in four component suspension data, there are significant additive baseline effects in the tecator data (supporting information, Figure S-2). Since the changes in physical properties of samples generally result in both additive baseline effects and multiplicative effects, the presence of significant additive baseline effects strongly suggests the existence of multiplicative effects. $OPLEC_m$ was therefore used to estimate the multiplicative parameter vector $\mathbf{p}$ for the calibration samples from the corresponding orthogonal projection pre-processed calibration spectra as described in section 3.3. During the estimation of the multiplicative parameter vector $\mathbf{p}$ for the calibration samples using $OPLEC_m$, the optimal number of columns included in $\mathbf{U}_s$ (i.e. r) is determined by scrutinizing the plot of $\min_{\mathbf{p}} f(\mathbf{p})$ versus r (Figure 4). It can be seen that $\min_{\mathbf{p}} f(\mathbf{p})$ drops sharply as the r increases from one to six, and then decreases slowly along with the further increase of r (Figure 4). One can therefore choose six as the optimal number of columns of $\mathbf{U}_s$.



294

295 Figure 4: The plot of $\min_{\mathbf{p}} f(\mathbf{p})$ versus the number of columns in $\mathbf{U}_s$ (i.e. r).

296 It is worth to point out again that the performance of $OPLEC_m$ is quite robust to the choice of *r* as long
297 as *r* is big enough but not too large. As shown in Figure 5, The RMSEP value of $OPLEC_m$ for the test
298 samples shows no significant difference when *r* taking a value between 6 and 11. In practice, such a
299 feature of $OPLEC_m$ can make it more user-friendly when being applied to complex systems.

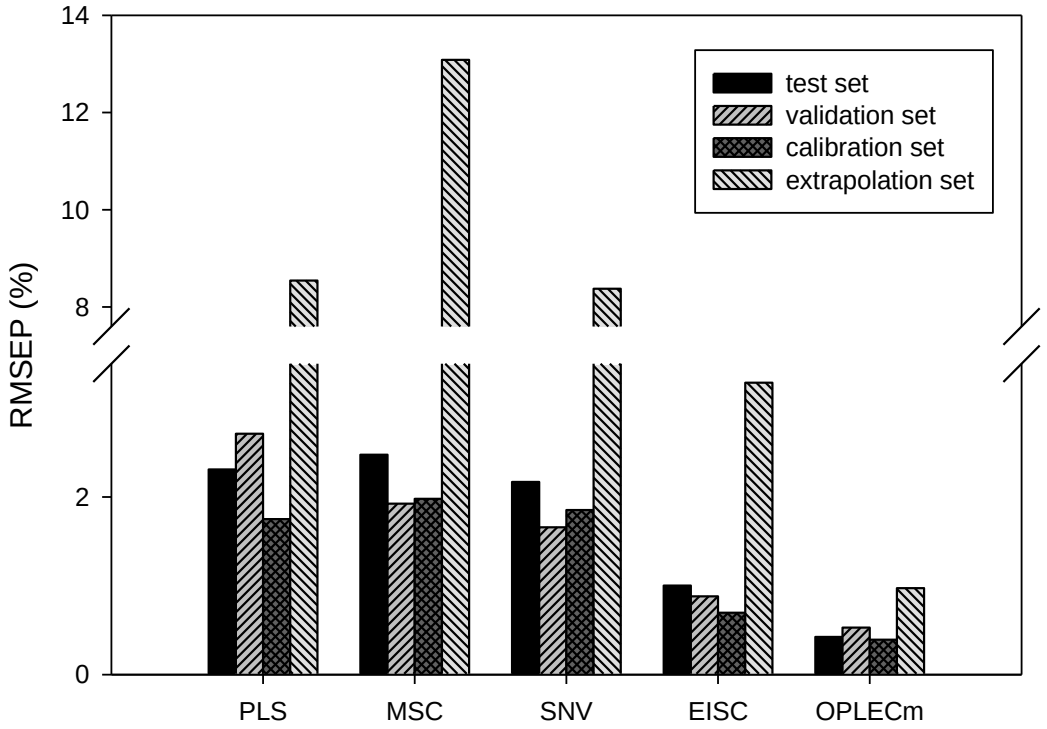


300

301 Figure 5: The RMSEP values for the test samples in the tecator data obtained by the optimal $OPLEC_m$
302 calibration models when $\mathbf{U}_s$ with different number of columns (i.e. r) were used in the calculation of the
303 multiplicative parameter vector **p** for the calibration spectra.

304

After the estimation of the multiplicative parameter vector **p** for the calibration samples, the dual calibration strategy of $OPLEC_m$ was adopted to mitigate the detrimental of multiplicative effects on the prediction of the fat content. PLS calibration models with and without the application of MSC, SNV and EISC were also established for comparison purposes. The optimal number of underlying components used in the dual calibration models of OPLEC as well as those PLS calibration models was chosen to be the one with minimal root-mean-square error of prediction (RMSEP) for the validation set. The results of $OPLEC_m$ along with those of the four optimal PLS calibration models with and without the application of MSC, SNV and EISC were shown in Figure 6.



313

314 Figure 6: The RMSEP values for the tecator data obtained by different calibration methods.

315 Figure 6 reveals that although the number of latent components (i.e. fourteen) used is sufficiently
316 large, the optimal PLS calibration model on the raw calibration spectra did not give satisfactory
317 predictions for all the four data sets. The RMSEP values for the calibration, validation, test and
318 extrapolation sets are 1.7%, 2.7%, 2.3% and 8.5%, respectively. The application of the empirical
319 multiplicative light scattering correction method, SNV saw no significant changes in the RMSEP values
320 for the four data sets. While preprocessing the spectral data by MSC resulted in a dramatic increase in
321 the RMSEP value for the extrapolation set which clearly demonstrates its limitation in practical
322 applications. The EISC preprocessing method surprisingly succeeded in improving the quality of the
323 predictions of PLS calibration model for the tecator data. Its RMSEP values for the calibration,
324 validation, test and extrapolation sets are 0.7%, 0.9%, 1.0% and 3.3%, respectively. The reasons of its
325 success in this particular data set are unclear. As expected, $OPLEC_m$ outperformed all the other methods
326 with RMSEP values for the calibration, validation, test and extrapolation sets equaling to 0.4%, 0.5%,
327 0.4% and 1.0%, respectively, This remarkable improvement further confirmed the effectiveness of
328 $OPLEC_m$ in mitigating the detrimental influence of multiplicative effects on the spectroscopic
329 quantitative analysis of heterogeneous mixture samples.
330
331
332 **5. Conclusion**
333 The separation of the spectral contributions due to variations in chemical compositions from
334 multiplicative effects caused by physical variations is crucial to the accurate quantitative analysis of
335 complex heterogeneous mixture samples using spectroscopic instruments. In this work, a modified

version of Optical Path-Length Correction and Estimation ($OPLEC_m$) method has been developed to correct the multiplicative effects in spectral measurements. $OPLEC_m$ differs from the original OPLEC method in the way of estimating the multiplicative parameters for the calibration samples. In $OPLEC_m$, the multiplicative parameters for the calibration samples were obtained by solving a constrained quadratic programming problem, which is much more efficient than the counterpart in the original OPLEC. Furthermore, a simple but effective method has been proposed for the determination of the model parameter involved (i.e. the number of spectroscopically active chemical components in the system under study). Due to the unique multiplicative parameter estimation strategy, the performance of $OPLEC_m$ is much more robust to the choice of the model parameter involved, which makes $OPLEC_m$ more user-friendly when being applied to complex systems. The performance of $OPLEC_m$ has been tested on four-component suspension spectral data set and one publicly available benchmark spectral data set. Experimental results reveal that $OPLEC_m$ can achieve satisfactory quantitative results from the spectroscopic measurements of heterogeneous mixtures. Compared with other existing methods designed for the same purpose, $OPLEC_m$ has features of implementation simplicity, wider applicability as well as better performance in terms of quantitative accuracy, and therefore has great potential in quantitative spectroscopic analysis of complex heterogeneous systems.

Acknowledgements

The authors acknowledge the financial support of the National Natural Science Foundation of China (grant no. 21075034), "973" National Key Basic Research Program of China (grant no. 2007CB310500) and the Fundamental Research Funds for the Central Universities of China and also Marie Curie FP6 (INTROSPECT)..

Supporting Information Available

MATLAB code for the modified OPLEC, the plots of $\mathbf{p}$ vs $\mathbf{U}_s\mathbf{U}_s^T\mathbf{p}$ and $diag(\mathbf{c}_1)\mathbf{p}$ vs $\mathbf{U}_s\mathbf{U}_s^T diag(\mathbf{c}_1)\mathbf{p}$ for the four component suspension data, the 129 raw calibration spectra of the tecator data. This material is available free of charge via the Internet at http://pubs.acs.org.

References：

(1) H.W. Siesler, Y. Ozaki, S. Kawata, H.M. Heise, Near-infrared spectroscopy: principal, instruments, applications,WILEY-VCH, Weinheim, **2002**

(2) P. Fayolle, D. Picque, G. Corrieu, Monitoring of fermentation processes producing lactic acid bacteria by mid-infrared spectroscopy, *Vib. Spectrosc.* **1997**, *14*, 247-252

(3) Y. Roggo, C. Roeseler, M. Ulmschneider, Near infrared spectroscopy for qualitative comparison of pharmaceutical batches, *J. Pharm. Biomed. Anal.* **2004**, *36*, 777–786

(4) A. Nordon, D. Littlejohn, A.S. Dann, P.A. Jeffkins, M.D. Richardson, S.L. Stimpson, In situ monitoring of a seed stage of a fermentation process using non-invasive NIR spectrometry, *The Analyst*, **2008**, *133*, 660-666

(5) Z.P. Chen, G. Fevotte, A. Caillet, D. Littlejohn, J. Morris, An advanced calibration strategy for in-situ quantitative monitoring of phase transition processes in suspensions using FT-Raman spectroscopy, *Anal. Chem.* **2008**, *80*, 6658-6665

(6) Z.P. Chen, J. Morris, A. Borissova, S. Khan, T. Mahmud, R. Penchev, K.J. Roberts, On-line monitoring of batch cooling crystallization of organic compounds using ATR-FTIR spectroscopy coupled with an advanced calibration method, *Chemom. Intell. Lab. Syst.* **2009**, *96*, 49–58

(7) P. Geladi, D. MacDougall, H. Martens, Linearization and Scatter-Correction for Near-Infrared Reflectance Spectra of Meat, *Appl. Spectrosc.* **1985**, *39 (3)*, 491-500

(8) I.A. Cowe, J.W. McNicol, The Use of Principal Components in the Analysis of Near-Infrared Spectra, *Appl. Spectrosc.* **1985**, *39 (2)*, 257-266

(9) H. Martens, M. Martens, Multivariate Analysis of Quality: An Introduction, John Wiley and Sons: Chichester, **2001**

(10) R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-trending of Near-Infrared Diffuse Reflectance Spectra, *Appl. Spectrosc.* **1989**, *43 (5)*, 772-777

(11) I.S. Helland, T. Næs, T. Isaksson, Related versions of the multiplicative scatter correction method for preprocessing spectroscopic data, *Chemom. Intell. Lab. Syst.* **1995**, *29 (2)*, 233-241

(12) D. Pedersen, H. Martens, J. Nielsen, S. Engelsen, Near-infrared absorption and scattering separated by extended inverted signal correction (EISC): Analysis of near-infrared transmittance spectra of single wheat seeds, *Appl. Spectrosc.* **2002**, *56 (9)*, 1206-1214

(13) H. Martens, J.P. Nielsen, S.B. Engelsen, Light Scattering and Light Absorbance Separated by Extended Multiplicative Signal Correction. Application to Near-Infrared Transmission Analysis of Powder Mixtures, *Anal. Chem.* **2003**, *75 (3)*, 394-404

(14) S.N. Thennadil, H. Martens, A. Kohler, Physics-based multiplicative scatter correction approaches for improving the performance of calibration models, *Appl. Spectrosc.* **2006**, *60*, 315-321

(15) R. Steponavicius, S.N. Thennadil, Extraction of chemical information of suspensions using radiative transfer theory to remove multiple scattering effects: application to a model two-component system, *Anal. Chem.* **2009**, *81*, 7713–7723

(16) R. Steponavicius, S.N. Thennadil, Extraction of chemical information of suspensions using Radiative transfer theory to remove multiple scattering effects: application to a model multicomponent system, *Anal. Chem.* **2011**, *83*, 1931-1937

(17) Z. Shi, C. Andersen, Pharmaceutical applications of separation of absorption and scattering in near-infrared spectroscopy (NIRS), *J. Pharm. Sci.* **2010**, *99*, 4766-4783

(18) W. Kessler, D. Oelkrug, R. Kessler, Using scattering and absorption spectra as MCR-hard model constraints for diffuse reflectance measurements of tablets, *Anal. Chim. Acta*, **2009**, *642*, 127–134

(19) Z.P.Chen, J. Morris, E. Martin, Extracting Chemical Information from Spectral Data with Multiplicative Light Scattering Effects by Optical Path-Length Estimation and Correction, *Anal. Chem.* **2006**, *78(9)*, 7674-7681

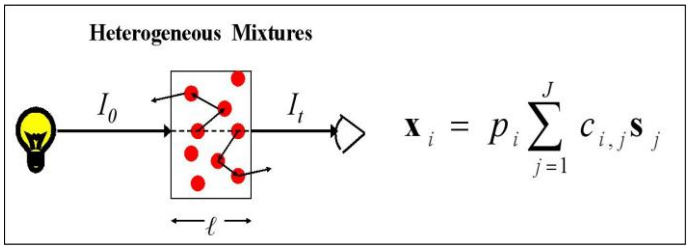
(20) Z.P.Chen, L.J. Zhong, A. Nordon, D. Littlejohn, M. Holden, M. Fazenda, L. Harvey, B. McNeil, J. Faulkner, J. Morris, Calibration of Multiplexed Fiber-Optic Spectroscopy, *Anal. Chem.* **2011**, *83(7)*, 2655-2659

(21) Z.P. Chen, J. Morris, Improving the linearity of spectroscopic data subjected to fluctuations in external variables by the extended loading space standardization, *The Analyst*, **2008**, *133*, 914-922

(22) C. Borggaard, H.H. Thodberg, Optimal minimal neural interpretation of spectra, *Anal. Chem.* **1992**, *64*, 545-551

424 **For TOC only**

425



426

Supporting Information

Title of the primary article:

Quantitative Spectroscopic Analysis of Heterogeneous Mixtures: the Correction of Multiplicative Effects Caused by Variations in Physical Properties of Samples

Authors' names:

Jing-Wen Jin[a], Zeng-Ping Chen[a], Li-Mei Li[a], Raimundas Steponavicius[b], Suresh N. Thennadil[c], Jing Yang[a] and Ru-Qin Yu*[a]*

Affiliations:

a. State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha, Hunan, 410082, PR China

b. School of Chemical Engineering and Advanced Materials, Newcastle University, Merz Court, Newcastle upon Tyne, NE1 7RU, United Kingdom

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

Table of content:

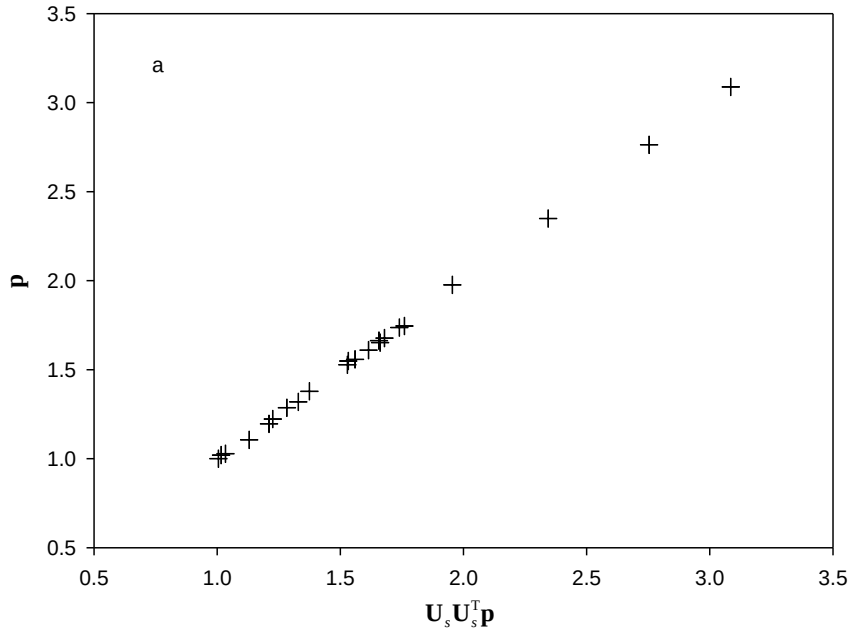
The MATLAB code for the modified OPLEC method

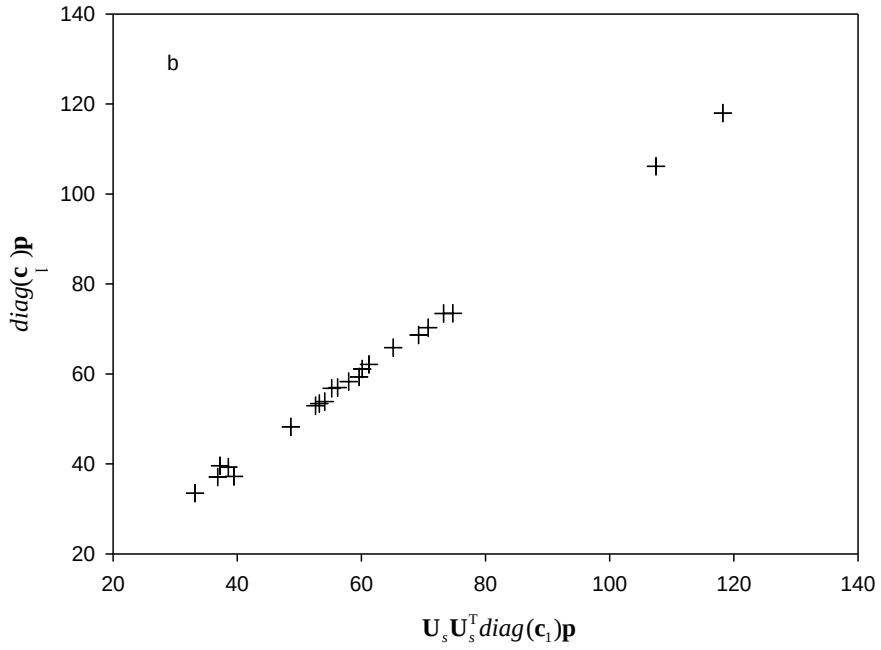
```
% [p, fval] = OPLEC_m(X, c, CompNumb);
% This is an m-file for the estimation of the multiplicative effect vector p for calibration samples;
% X contains x_i in its rows; x_i (i = 1,2,···,I) are the spectra of I calibration samples.
% c is the concentration vector of the target chemical component in the calibration samples;
% CompNumb is the number of spectroscopically active chemical components in mixture samples;
% p is a vector containing the multiplicative scattering parameters for the calibration samples;
% fval is the value of objective function at p;

function [p, fval]=OPLECm(X, c, CompNumb);
[U,S,V]=svd(X);
Us= U(:,1:CompNumb);
n=length(c);
w=max(c);
H1=eye(n, n)- Us* Us';
H2= diag(c./w)*H1* diag(c./w);
```

```
H=H1+H2;    % matrix H in min(0.5*p'*H*p+f'*p);
f=zeros(n,1);    % vector f in min(0.5*p'*H*p+f'*p);
A=-eye(n,n);    % matrix A in A*p<=b;
b=-ones(n,1);    % vector b in A*p<=b;
StartingVect=ones(n,1);
options=optimset('quadprog');
options=optimset(options,'LargeScale','off','Display','off');
[p,fval]=quadprog(H,f,A,b,[],[],[],[],StartingVect,options);
```

% After obtaining the model parameter vector **p** for calibration samples, two calibration models are built using the standard PLS toolbox. One is between the concentration vector (**c**) of the target chemical component and the spectral data **X;** the other is between $diag(\mathbf{c})\mathbf{p}$ and **X**. The multiplicative effect on the test sample can then be corrected through dividing the prediction of the second calibration model by the prediction of the first calibration model.

1) **Figure S-1: The plots of p vs $\mathbf{U}_s\mathbf{U}_s^T\mathbf{p}$ (a) and $diag(\mathbf{c}_1)\mathbf{p}$ vs $\mathbf{U}_s\mathbf{U}_s^T diag(\mathbf{c}_1)\mathbf{p}$ (b) for the four component suspension data. The number of columns in $\mathbf{U}_s$ is three.**





2) Figure S-2: The 129 raw calibration spectra of the tecator data.