analytical chemistry

ARTICLE

pubs.acs.org/ac

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

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

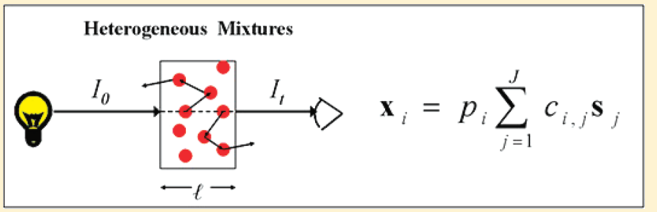
[†]State Key Laboratory of Chemo/Biosensing and Chemometrics, College of Chemistry and Chemical Engineering, Hunan University, Changsha, 410082, China

[‡]School of Chemical Engineering and Advanced Materials, Newcastle University, Merz Court, Newcastle upon Tyne, NE1 7RU, United Kingdom

[§]Chemical and Process Engineering, University of Strathclyde, 75 Montrose Street, Glasgow, G1 1XJ, United Kingdom

Supporting Information

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 components in the same sample mixture. On the basis of 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.



The quantitative analysis of heterogeneous mixture samples using conventional instruments such as high-performance liquid chromatography (HPLC) generally involves troublesome and time-consuming sample preparations. Because of their high measuring speed, multiplicity of analysis, nondestructivity, flexibility, and especially the requirement of less or even no sample preparation, spectroscopic technologies such as near-infrared (NIR), mid-infrared (MIR), and Raman spectroscopy 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 principal component regression (PCR)[8] and partial least-squares (PLS)[9] which postulate 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

Received: September 30, 2011
Accepted: November 16, 2011
Published: November 16, 2011

dx.doi.org/10.1021/ac202598f | *Anal. Chem.* 2012, 84, 320–326

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 preprocessing 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, Steponavicius and Thennadil 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 NIR 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 the Kubelka−Munk equation with multivariate curve resolution (MCR).[18] Like the method based on radiative transfer theory, the application of the 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 simple but effective approach for determining the optimal model parameter (i.e., the number of spectroscopically active chemical components) in OPLEC, (3) to improve the robustness of OPLEC when being applied to complex systems, and finally (4) to evaluate the performance of the modified OPLEC method on two publicly available benchmark data sets.

THEORY

Dual-Calibration Strategy Adopted by OPLEC to Correct Multiplicative Effects. For spectral measurements with multiplicative effects caused by changes in the optical path length due to the physical variations of the samples, the measured spectrum ($\mathbf{x}_i$, row vector) of sample i composed of J chemical components can be approximated by the following model:[6,7,21]

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

where $c_{i,j}$ is the concentration of the jth chemical component in the ith mixture sample; $\mathbf{s}_j$ represents the pure spectrum of the jth chemical component in the mixtures. The coefficient p_i accounts for the multiplicative effects in the spectral measurements of the ith sample caused by changes in the optical path length due to the physical variations of the sample; I denotes the number of calibration samples. Assume the first component is the target constituent in the mixtures and $\sum_{j=1}^{J} c_{i,j} = 1$ (which strictly hold for $c_{i,j}$ representing unit-free concentration such as weight fraction and mole fraction), then eq 1 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 the Multiplicative Parameter Estimation section), 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 $\mathrm{diag}(\mathbf{c}_1)\mathbf{p}$ ($\mathrm{diag}(\mathbf{c}_1)\mathbf{p} = [p_1 \times c_{1,1}; p_2 \times c_{2,1}; \ldots; p_I \times c_{I,1}]$). 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.

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

dx.doi.org/10.1021/ac202598f |Anal. Chem. 2012, 84, 320–326

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 $\mathbf{p}$ for the calibration samples was proposed in this work.

Suppose the singular value decomposition of $\mathbf{X}$ ($\mathbf{X} = [\mathbf{x}_1; \mathbf{x}_2; ...; \mathbf{x}_I]$) 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, $\mathbf{E} = \mathbf{U}_n\Sigma_n\mathbf{V}_n^{\mathrm{T}}$; 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 $\mathbf{p}$ and $\mathrm{diag}(\mathbf{c}_1)\mathbf{p}$ are in the column space of $\mathbf{U}_s$, so the following equations hold:

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

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

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

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

where $\|\ \|_2$ denotes the 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 the Supporting Information):

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

Determination of the Number of Columns in $\mathbf{U}_s$. 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 a 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.

CASE STUDIES

The effectiveness of the modified OPLEC method (hereafter referred to $\mathrm{OPLEC_m}$) with respect to its ability to estimate multiplicative parameters was first tested on the NIR total diffuse transmittance spectra of a four-component suspension system consisting of water, deuterium, ethanol, and polystyrene (hereafter referred to as four-component suspension data). To further explore the potential of $\mathrm{OPLEC_m}$, another real-world NIR transmittance spectra of meat samples recorded on a Tecator Infratec food and feed analyzer (hereafter referred to as tecator data) is employed. This spectral data set is publicly available and hence ensures that the interested reader can repeat the analysis.

Four-Component Suspension Data[16]. The four-component suspension system is composed of three fully miscible absorbing species of water, deuterium oxide, ethanol and a species that both absorbs and scatters light (i.e., a particulate species of polystyrene). Specifically, the ranges 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 2 nm, 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 20 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]

Tecator Data[22]. This benchmark spectral data set consists of the NIR absorbance spectra of 240 meat samples recorded on a Tecator Infratec food and feed analyzer working in the wavelength range of 850−1050 nm with an interval of 2 nm by the NIR transmission principle. Each sample contains finely chopped pure meat with different moisture, fat, and protein contents. A Soxhlet method was used as the laboratory reference for fat determination. The Soxhlet values ranged from 2% to 59% fat. The 240 spectra were divided into five data sets for the purpose of model validation and extrapolation studies (calibration set, 129; validation set, 43; test set, 43; extrapolation set for fat, 8; extrapolation set for protein, 7). The task in the present work is restricted to predict the fat content (%) of a meat sample on the basis of its NIR absorbance spectrum; the extrapolation set for

dx.doi.org/10.1021/ac202598f |*Anal. Chem.* 2012, 84, 320–326

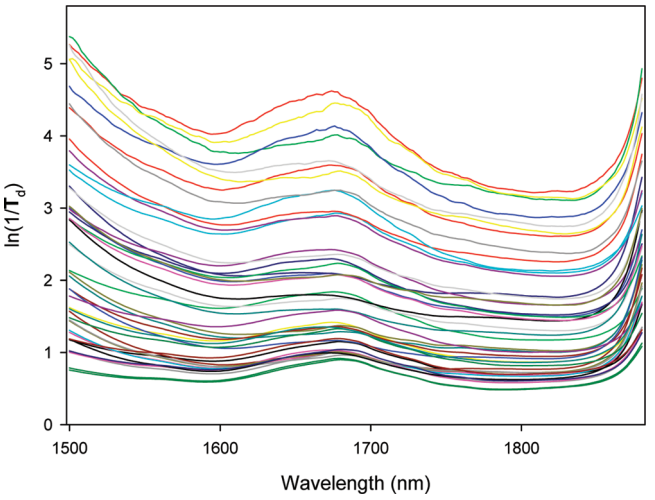


Figure 1. Raw spectra of the four-component suspension system.

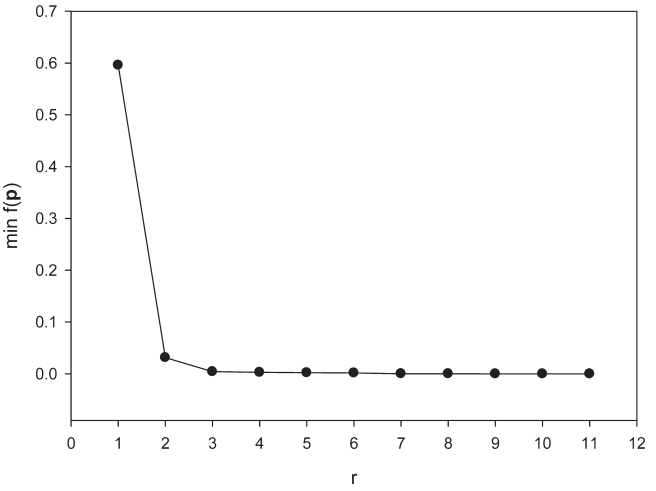


Figure 2. Relationship between $\min_p f(\mathbf{p})$ and the number of columns of $\mathbf{U}_s$ (i.e., r) for the four-component suspension data.

protein is therefore excluded. The tecator data is available at http://lib.stat.cmu.edu/data sets/tecator.

Data Pretreatment. For the aforementioned two data sets, the possible additive baseline effects and wavelength-dependent spectral variations were first removed by projecting the measured spectra onto the orthogonal complement of the space spanned by the row vectors of $\mathbf{M} = [\mathbf{1}; \boldsymbol{\lambda}; \boldsymbol{\lambda}^2]$.[19] The preprocessed spectra were then used to calculate the multiplicative parameter vector $\mathbf{p}$ for the calibration samples. The dual-calibration models in OPLEC$_m$ were built on the preprocessed spectra by using the PLS method. The predictive performance of OPLEC$_m$ was compared with those of PLS calibration models with and without the application of data preprocessing methods such as MSC, SNV, EISC, and EMSC as long as they are applicable. The root-mean-square error of prediction (RMSEP) was used to assess the performance of the calibration models.

RESULTS AND DISCUSSION

Four-Component Suspension Data. The raw total transmittance spectra of the four-component suspension samples are presented in Figure 1. It can be observed that the variations in polystyrene particle size and concentration across samples resulted in significant additive baseline shift as well as multiplicative effects in the spectral data. Though the additive baseline effects and possible wavelength-dependent spectral variations can be readily removed by orthogonal projection preprocessing, the multiplicative effects as a consequence of the changes in sample's effective optical path length are rather difficult to correct. Such multiplicative effects cannot be effectively modeled by multivariate linear calibration models either. Without being properly corrected or modeled, they can significantly deteriorate the predictive performance of multivariate linear calibration models.[13,19]

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 preprocessed 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 $\min_p f(\mathbf{p})$ versus r (Figure 2). From Figure 2, it can be seen that $\min_p f(\mathbf{p})$ 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 $\min_p f(\mathbf{p})$, which means the most spectral information relevant to $\mathbf{p}$ and $\text{diag}(\mathbf{c}_1)\mathbf{p}$ was included in the first three principal components of $\mathbf{U}_s$. Therefore, the optimal value of r was then set to three.

After the estimation of the multiplicative parameter vector $\mathbf{p}$ for the calibration samples, one can assess the applicability of OPLEC$_m$ to the spectral data set by examining the two plots of $\mathbf{p}$ versus $\mathbf{U}_s\mathbf{U}_s^T\mathbf{p}$ and $\text{diag}(\mathbf{c}_1)\mathbf{p}$ versus $\mathbf{U}_s\mathbf{U}_s^T\,\text{diag}(\mathbf{c}_1)\mathbf{p}$, respectively (Supporting Information, Figure S-1). As shown in Supporting Information Figure S-1, both $\mathbf{p}$ and $\text{diag}(\mathbf{c}_1)\mathbf{p}$ are in good agreement with $\mathbf{U}_s\mathbf{U}_s^T\mathbf{p}$ and $\mathbf{U}_s\mathbf{U}_s^T\,\text{diag}(\mathbf{c}_1)\mathbf{p}$, 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. Supporting Information 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 compares 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 cannot be fully corrected by MSC, SNV, or EISC. Hence, the predictive errors of the PLS calibration models built on the calibration spectra preprocessed 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 preprocessed methods. The OPLEC$_m$ calibration model with

dx.doi.org/10.1021/ac202598f |*Anal. Chem.* 2012, 84, 320–326

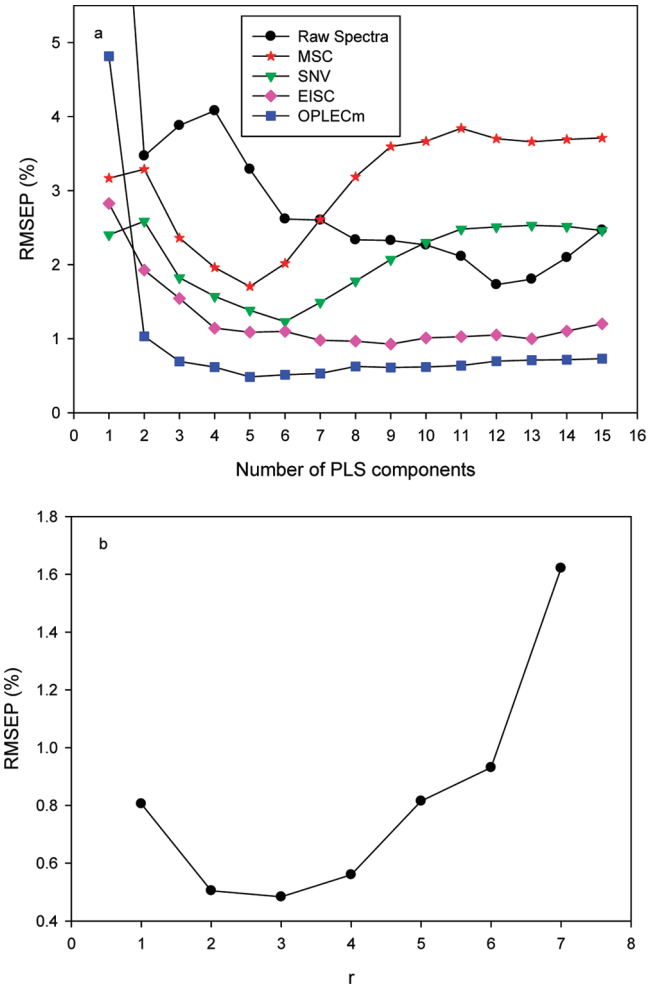


Figure 3. (a) Predictive performance of $OPLEC_m$ and the PLS models built on the calibration spectra of the four-component suspension system preprocessed by different methods (black circle, the raw spectra; red star, MSC; green triangle down, SNV; pink diamond, EISC; blue square, $OPLEC_m$). (b) The 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.

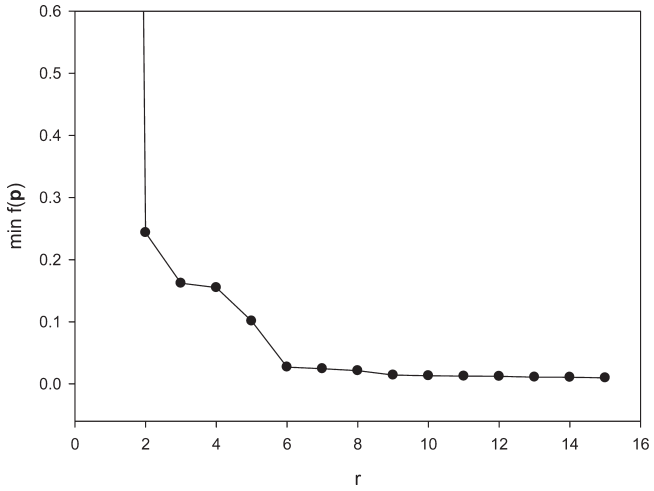


Figure 4. Plot of $\min_p f(\mathbf{p})$ vs the number of columns in $\mathbf{U}_s$ (i.e., r).

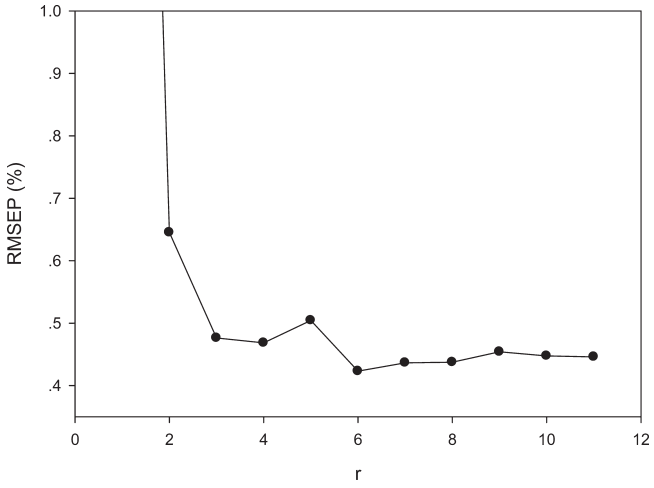


Figure 5. RMSEP values for the test samples in the tecator data obtained by the optimal $OPLEC_m$ calibration models when $\mathbf{U}_s$ with different number of columns (i.e., r) were used in the calculation of the multiplicative parameter vector $\mathbf{p}$ for the calibration spectra.

five underlying components provided the best predictive results with an $RMSEP_{test}$ value as low as 0.005, whereas the corresponding best $RMSEP_{test}$ value of the PLS calibration model with nine underlying components on the calibration spectra preprocessed 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.

Tecator Data. As in the 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 preprocessed calibration spectra as described in the Data Pretreatment section. 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 $\min_p f(\mathbf{p})$ versus r (Figure 4). It can be seen that $\min_p f(\mathbf{p})$ 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 $\mathbf{U}_s$.

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

After the estimation of the multiplicative parameter vector $\mathbf{p}$ 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,

dx.doi.org/10.1021/ac202598f |*Anal. Chem.* 2012, 84, 320–326

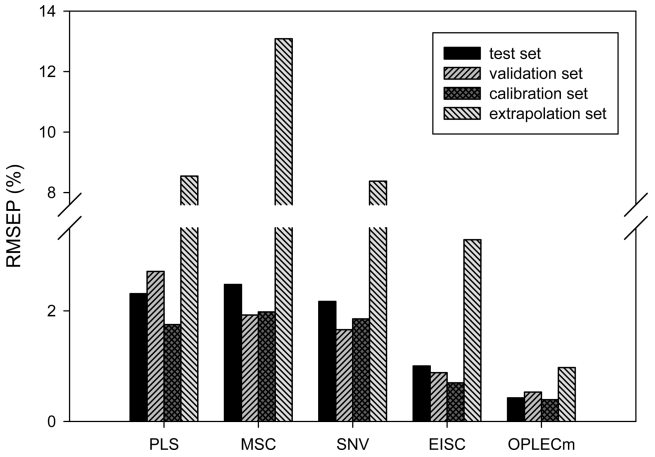


Figure 6. RMSEP values for the tecator data obtained by different calibration methods.

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 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 are shown in Figure 6.

Figure 6 reveals that, although the number of latent components (i.e., 14) used is sufficiently large, the optimal PLS calibration model on the raw calibration spectra did not give satisfactory predictions for all the four data sets. The RMSEP values for the calibration, validation, test, and extrapolation sets are 1.7%, 2.7%, 2.3%, and 8.5%, respectively. The application of the empirical multiplicative light scattering correction method, SNV, saw no significant changes in the RMSEP values for the four data sets, whereas preprocessing the spectral data by MSC resulted in a dramatic increase in the RMSEP value for the extrapolation set which clearly demonstrates its limitation in practical applications. The EISC preprocessing method surprisingly succeeded in improving the quality of the predictions of PLS calibration model for the tecator data. Its RMSEP values for the calibration, validation, test, and extrapolation sets are 0.7%, 0.9%, 1.0%, and 3.3%, respectively. The reasons of its success in this particular data set are unclear. As expected, $OPLEC_m$ outperformed all the other methods with RMSEP values for the calibration, validation, test, and extrapolation sets equaling to 0.4%, 0.5%, 0.4%, and 1.0%, respectively, This remarkable improvement further confirmed the effectiveness of $OPLEC_m$ in mitigating the detrimental influence of multiplicative effects on the spectroscopic quantitative analysis of heterogeneous mixture samples.

CONCLUSION

The separation of the spectral contributions due to variations in chemical compositions from multiplicative effects caused by physical variations is crucial to the accurate quantitative analysis of complex heterogeneous mixture samples using spectroscopic instruments. In this work, a modified version of the 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 a 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 and 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.

ASSOCIATED CONTENT

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

AUTHOR INFORMATION

Corresponding Author

*Phone: (+86) 731 88821916. Fax: (+86) 731 88821916. E-mail: zpchen2002@hotmail.com (Z.-P.C.); rqyu@hnu.cn (R.-Q.Y.).

ACKNOWLEDGMENT

The authors acknowledge the financial support of the National Natural Science Foundation of China (Grant No. 21075034 and Grant No. 21035001), the National Instrumentation Program of China (Grant No. 2011YQ03012401), and the Fundamental Research Funds for the Central Universities of China, and also Marie Curie FP6 (INTROSPECT).

REFERENCES

(1) Siesler, H. W.; Ozaki, Y.; Kawata, S.; Heise, H. M. *Near-Infrared Spectroscopy: Principal, Instruments, Applications*; Wiley-VCH: Weinheim, Germany, 2002.

(2) Fayolle, P.; Picque, D.; Corrieu, G. *Vib. Spectrosc.* **1997**, *14*, 247–252.

(3) Roggo, Y.; Roeseler, C.; Ulmschneider, M. *J. Pharm. Biomed. Anal.* **2004**, *36*, 777–786.

(4) Nordon, A.; Littlejohn, D.; Dann, A. S.; Jeffkins, P. A.; Richardson, M. D.; Stimpson, S. L. *Analyst* **2008**, *133*, 660–666.

(5) Chen, Z. P.; Fevotte, G.; Caillet, A.; Littlejohn, D.; Morris, J. *Anal. Chem.* **2008**, *80*, 6658–6665.

(6) Chen, Z. P.; Morris, J.; Borissova, A.; Khan, S.; Mahmud, T.; Penchev, R.; Roberts, K. J. *Chemom. Intell. Lab. Syst.* **2009**, *96*, 49–58.

(7) Geladi, P.; MacDougall, D.; Martens, H. *Appl. Spectrosc.* **1985**, *39* (3), 491–500.

(8) Cowe, I. A.; McNicol, J. W. *Appl. Spectrosc.* **1985**, *39* (2), 257–266.

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

(10) Barnes, R. J.; Dhanoa, M. S.; Lister, S. J. *Appl. Spectrosc.* **1989**, *43* (5), 772–777.
(11) Helland, I. S.; Næs, T.; Isaksson, T. *Chemom. Intell. Lab. Syst.* **1995**, *29* (2), 233–241.
(12) Pedersen, D.; Martens, H.; Nielsen, J.; Engelsen, S. *Appl. Spectrosc.* **2002**, *56* (9), 1206–1214.
(13) Martens, H.; Nielsen, J. P.; Engelsen, S. B. *Anal. Chem.* **2003**, *75* (3), 394–404.
(14) Thennadil, S. N.; Martens, H.; Kohler, A. *Appl. Spectrosc.* **2006**, *60*, 315–321.
(15) Steponavicius, R.; Thennadil, S. N. *Anal. Chem.* **2009**, *81*, 7713–7723.
(16) Steponavicius, R.; Thennadil, S. N. *Anal. Chem.* **2011**, *83*, 1931–1937.
(17) Shi, Z.; Andersen, C. *J. Pharm. Sci.* **2010**, *99*, 4766–4783.
(18) Kessler, W.; Oelkrug, D.; Kessler, R. *Anal. Chim. Acta* **2009**, *642*, 127–134.
(19) Chen, Z. P.; Morris, J.; Martin, E. *Anal. Chem.* **2006**, *78* (9), 7674–7681.
(20) Chen, Z. P.; Zhong, L. J.; Nordon, A.; Littlejohn, D.; Holden, M.; Fazenda, M.; Harvey, L.; McNeil, B.; Faulkner, J.; Morris, J. *Anal. Chem.* **2011**, *83* (7), 2655–2659.
(21) Chen, Z. P.; Morris, J. *Analyst* **2008**, *133*, 914–922.
(22) Borggaard, C.; Thodberg, H. H. *Anal. Chem.* **1992**, *64*, 545–551.