Multivariate calibration of near infrared spectroscopy in the presence of light scattering effect: a comparative study

Ke Wang, Guoyi Chi, Raymond Lau, Tao Chen *

School of Chemical and Biomedical Engineering, Nanyang Technological University, 62 Nanyang Drive, Singapore 637459, Singapore

Abstract:

When analyzing heterogeneous samples using spectroscopy, the light scattering effect introduces non-linearity into the measurements and deteriorates the prediction accuracy of conventional linear models. This paper compares the prediction performance of two categories of chemometric methods: pre-processing techniques to remove the non-linearity, and non-linear calibration techniques to directly model the non-linearity. A rigorous statistical procedure is adopted to ensure reliable comparison. The results suggest that optical path length estimation and correction (OPLEC) and Gaussian process (GP) regression are the most promising among the investigated methods. Furthermore, the combination of pre-processing and non-linear models is explored with limited success being achieved.

Keywords: Light scattering correction; Non-linear multivariate calibration; Pre-processing.

1. Introduction

Recently, the application of near infrared (NIR) spectroscopy has attracted increasing attention in pharmaceutical, petrochemical and food industries due to its rapid, powerful and economical analysis of a large number of samples. However, when NIR is used to analyze suspension or powder, the differences in particle size and shape, sample packing and sample surface (and other properties) can result in sample-to-sample variation in the optical path length, giving rise to the well-known light scattering effect (Geladi, Macdougall, and Martens 1985). According to Beer-Lambert's law, the absorbance spectra are linear to the analyte properties (e.g. concentration). However, the light scattering can introduce uncertain non-linear variation to the spectral data (Geladi, Macdougall, and Martens 1985). As a result, the traditional linear calibration methods (such as principal component regression (PCR) and partial least squares (PLS)) typically give inferior prediction results.

Several strategies have been proposed to account for the light scattering effect for improving calibration modelling. One possibility is to improve the spectrometers by using advanced *hardware* configuration based on the principle of the light propagation (Gratton 1971). Alternatively, the light scattering effect can be

* Corresponding author. Tel.: +65 6513 8267; Fax: +65 6794 7553 (T. Chen).

Email addresses: chentao@ntu.edu.sg (T. Chen).

corrected or compensated using chemometric (*software*) techniques (Martens, Nielsen, and Engelsen 2003). In principle, combining the above hardware and software strategies is expected to give better results than either (Thennadil, Martens, and Kohler 2006). Hardware solutions appear to be desired since they are based on fundamental physical principles. However, they may not be compatible with the spectroscopic instruments currently used in practice, and thus significantly extra cost is incurred. Therefore, in this study, we focus on chemometric methods to deal with the light scattering effect for calibration modelling, which is more cost-effective for existing instruments.

The various chemometric techniques proposed in the literature to handle light scattering can be classified into two categories. One category is the pre-processing methods to remove the non-linear variation of the data prior to applying PLS, such as first (D1) and second derivatives (D2) (Savitzky and Golay 1964; Brown, Vega-Montoto, and Wentzell 2000), standard normal variate (SNV) (Barnes, Dhanoa, and Lister 1989), extended multiplicative signal correction (EMSC) (Martens, Nielsen, and Engelsen 2003), extended inverted signal correction (EISC) (Pedersen et al. 2002), and optical path length estimation and correction (OPLEC) (Chen, Morris, and Martin 2006). The other strategy is to directly model the non-linearity in the data, such as artificial neural network (ANN) (Li et al. 2001; Long, Gregoriou, and Gemperline 1990; Wang, Hwang, and Kowalski 1995; Thodberg 1996), least squares support vector machine (LS-SVM) (Suykens et al. 2002; Thissen et al. 2004) and Gaussian process (GP) (Chen, Morris, and Martin 2007). As a consequence of the rapid development of scattering-corrected calibration models, rigorous assessment and comparison of available techniques are required, and they form the basis of this paper. In previous comparative studies, some chemometric methods, such as D1, D2, SVN, EMSC, EISC and ANN, have been investigated (Hadjjiiski, Geladi, and Hopke 1999; Centner et al. 2000; Thissen et al. 2004; Thennadil and Martin 2005; Balabin, Safieva, and Lomakina 2007). However, several novel techniques, including OPLEC, LS-SVM and GP, have emerged in recent years with very promising results being achieved. Therefore, our study serves an update of the new development in this field. We are particularly interested in the relative advantage of pre-processing methods and non-linear calibration models, since they originate from different rationales. Furthermore, we will explore the option of combining pre-processing and non-linear calibration methods in an attempt to achieve more accurate prediction than individual techniques.

This comparative study is based on three NIR datasets that are publicly available from the Internet. The predictive accuracy, quantified by root mean square error of prediction (RMSEP), is used to evaluate the performance of these investigated calibration methods. A rigorous statistical procedure is adopted to ensure that the comparison results are meaningful and reliable. Firstly, for developing calibration models, each dataset needs to be divided into training and testing data. It is crucial to ensure that the training/testing data are representative of the whole dataset in order to obtain reliable and robust evaluation results. This objective will be achieved by a random splitting strategy with 50 repeats. Secondly, a statistical hypothesis test approach, the paired *t*-test, will be used to determine whether one method is significantly (in a rigorous statistical sense) better than the other. More details will be discussed in Section 3.2.

2. Theory

This section briefly presents the theory of the pre-processing techniques (SNV, EMSC, EISC, OPLEC) and the non-linear calibration methods (ANN, GP, LS-SVM). Throughout this section, x_{ij} denotes the measured spectra of the i^{th} ($i = 1, \dots, I$) sample at the j^{th} ($j = 1, \dots, J$) wavelength (variable), and $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{iJ}]$; y_i represents the corresponding response variable (such as concentration) for the i^{th} sample.

2.1. Pre-processing methods

Light scattering effect brings in uncertain non-linear variations in the data, which adversely affects the prediction accuracy of the conventional linear calibration models. The pre-processing methods aim at removing the non-linearity in the data, and then linear models can be applied.

2.1.1. Standard normal variate

The model of standard normal variate (SNV) can be written as $\mathbf{x}_i = a_i + b_i \mathbf{x}_{i,chem}$, where $\mathbf{x}_{i,chem}$ denotes the theoretical spectra. The coefficients, a_i and b_i , are the mean and the standard deviation of the i^{th} sample with respect to wavelengths. Therefore, the corrected spectra are obtained by $\mathbf{x}_{i,corr} = (\mathbf{x}_i - a_i)/b_i$.

2.1.2. Multiplicative signal correction and its extension

The model of multiplicative signal correction (MSC) is similar to SNV as $\mathbf{x}_i = a_i + b_i \mathbf{x}_{i,chem}$. In MSC, it is assumed $\mathbf{x}_{i,chem} \approx \mathbf{m}$, where $\mathbf{m}$ denotes the reference spectra. The coefficients, a_i and b_i , can be estimated by least squares regression of $\mathbf{x}_i$ to $\mathbf{m}$ (Cornel, Lindenberg, and Mazzotti 2008). However, MSC is restricted by the assumption that the chemical variation between the theoretical and reference spectra can be ignored. Otherwise, the estimated coefficients may contain certain irrelevant information (Chen, Morris, and Martin 2006). Recently, extended multiplicative signal correction (EMSC) was developed to address this issue (Martens, Nielsen, and Engelsen 2003). The model of EMSC is given by $\mathbf{x}_i = a_i + b_i \mathbf{x}_{i,chem} + d_i \boldsymbol{\lambda} + e_i \boldsymbol{\lambda}^2$, where $\boldsymbol{\lambda}$ is the wavelength vector; d_i and e_i represent the smooth wavelength-dependent spectral variations. According to Beer-Lambert's law, the theoretical spectra $\mathbf{x}_{i,chem}$ can be expressed as $\mathbf{x}_{i,chem} = \sum_{k=1}^K y_{ik} \mathbf{s}_k$, where y_{ik} denotes the concentration of k^{th} constituent in i^{th} sample and $\mathbf{s}_k$ is the pure absorption spectra of k^{th} constituent. When the pure spectra are available, the model parameters can be estimated using least squares and the corrected spectra are given by $\mathbf{x}_{i,corr} = (\mathbf{x}_i - a_i - d_i \boldsymbol{\lambda} - e_i \boldsymbol{\lambda}^2)/b_i$. When this is not possible, the mean spectra of entire dataset are typically used in place of $\mathbf{x}_{i,chem}$ for parameter estimation.

2.1.3. Inverted signal correction and its extension

Inverted signal correction (ISC) is based on the "reverse" expression of MSC as $\mathbf{m} = a_i + b_i \mathbf{x}_i$ (Pedersen et al. 2002). The parameters a_i and b_i are estimated by least squares regression of $\mathbf{m}$ to $\mathbf{x}_i$. Hence the corrected spectra are written as $\mathbf{x}_{i,corr} = a_i + b_i \mathbf{x}_i$. As MSC has been extended to EMSC, ISC can be also extended to EISC as $\mathbf{m} = a_i + b_i \mathbf{x}_i + c_i \mathbf{x}_i^2 + d_i \boldsymbol{\lambda} + e_i \boldsymbol{\lambda}^2$. The parameters a_i , b_i , c_i , d_i , e_i can be estimated by least squares, and the corrected spectra are given by $\mathbf{x}_{i,corr} = a_i + b_i \mathbf{x}_i + c_i \mathbf{x}_i^2 + d_i \boldsymbol{\lambda} + e_i \boldsymbol{\lambda}^2$.

2.1.4. Optical path length estimation and correction

Optical path length estimation and correction (OPLEC) was devised using the same model as EMSC of $\mathbf{x}_i = a_i + b_i \mathbf{x}_{i,chem} + d_i \lambda + e_i \lambda^2$ (Chen, Morris, and Martin 2006), the difference being the parameter estimation algorithms. The parameters a_i , d_i and e_i can be removed by projecting $\mathbf{x}_i$ onto the orthogonal complement of the space spanned by $\mathbf{P} = [\mathbf{1}; \lambda; \lambda^2]$ as $\mathbf{z}_i = \mathbf{x}_i(\mathbf{I} - \mathbf{P}^+ \mathbf{P}) = (a_i + b_i \mathbf{x}_{i,chem} + d_i \lambda + e_i \lambda^2)(\mathbf{I} - \mathbf{P}^+ \mathbf{P}) = \sum_{k=1}^K b_i y_{ik} \mathbf{p}_k$. Here $\mathbf{z}_i$ is the standardized spectra, $\mathbf{p}_k = \mathbf{s}_k(\mathbf{I} - \mathbf{P}^+ \mathbf{P})$, $\mathbf{I}$ is an identity matrix with the same dimension as $\mathbf{P}^+ \mathbf{P}$, and $\mathbf{P}^+$ represents the pseudo-inverse of $\mathbf{P}$. The parameter $\mathbf{b} = [b_1; \dots b_I]$ could be obtained by a partitioning strategy which was detailed in (Chen, Morris, and Martin 2006). Having obtained $\mathbf{b}$, two new regression vectors, β_1 and β_2 , can be estimated by the following two models of $\text{diag}(\mathbf{y}_1) \mathbf{b} = [\mathbf{1}, \mathbf{Z}] \beta_1$ and $\mathbf{b} = [\mathbf{1}, \mathbf{Z}] \beta_2$ through a multivariate linear regression (e.g. PLS). Then β_1 and β_2 are used to predict the concentration of a test sample as $y_{\text{test}} = ([\mathbf{1}, \mathbf{z}_{\text{test}}] \beta_1) / ([\mathbf{1}, \mathbf{z}_{\text{test}}] \beta_2)$.

2.2. Non-linear calibration methods

Since the relationship between the measured spectra and the response variable is non-linear due to light scattering effect, the use of the non-linear regression models may be a natural choice for calibration. This subsection gives an overview of the state-of-the-art non-linear calibration methods of ANN, GP and LS-SVM.

2.2.1. Artificial neural networks

A typical feed-forward artificial neural network (ANN) consists of three layers (input, hidden and output layer), each layer comprising multiple neurons. In the calibration modelling, the response y_i from the output layer can be expressed as (Wang, Hwang, and Kowalski 1995): $y_i = f[\sum_{n=1}^N \omega_n g(\sum_{j=1}^J (w_j x_{ij} + \phi_j)) + \varepsilon_n]$, where N is the number of hidden-layer neurons, w_j represents the weights connecting the input- and hidden-layer neurons, ω_n indicates the weights connecting the hidden- and output-layer neurons, ϕ_j and ε_n are the biases in the hidden and output layers. The “transfer functions”, $f(\cdot)$ and $g(\cdot)$, typically are taken as linear and sigmoid functions. The parameters are estimated by the “back-propagation” algorithm (Bishop 1995). In this study, we adopt a Bayesian back-propagation algorithm based on a Gaussian approximation to the posterior distribution of parameters (MacKay 1992).

2.2.2. Gaussian process regression

In a GP regression model (Chen, Morris, and Martin 2007), the response variable y_i is modelled by a joint Gaussian distribution with zero mean as $\mathbf{y} = (y_1, \dots, y_I) \sim G(\mathbf{0}, \mathbf{C})$. Here $\mathbf{C}$ is an $I \times I$ covariance matrix

whose elements can be defined by $C_{ik} = C(\mathbf{x}_i, \mathbf{x}_k) = a_0 + a_1 \sum_{j=1}^J x_{ij} x_{kj} + v_0 \exp\left(-\sum_{j=1}^J w_j (x_{ij} - x_{kj})^2\right) + \sigma^2 \delta_{ik}$, where the first two terms represent the constant bias and linear correlation, the third term is

similar to the form of the radial basis function and the fourth term corresponds to the random error. The model parameters, denoted by $\boldsymbol{\theta} = (a_0, a_1, v_0, w_1, \dots, w_J, \sigma^2)$, can be estimated by maximizing the log-likelihood of

$\log p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{1}{2} \log |\mathbf{C}| - \frac{1}{2} \mathbf{y}^T \mathbf{C}^{-1} \mathbf{y} - \frac{I}{2} \log(2\pi)$. For a new data point $\mathbf{x}^*$, the prediction y^* is also

Gaussian distributed with mean $E(y^*) = \mathbf{k}^T(\mathbf{x}^*)\mathbf{C}^{-1}\mathbf{y}$ and variance $\text{Var}(y^*) = C(\mathbf{x}^*, \mathbf{x}^*) - \mathbf{k}^T(\mathbf{x}^*)\mathbf{C}^{-1}\mathbf{k}(\mathbf{x}^*)$, where $\mathbf{k}(\mathbf{x}^*) = [C(\mathbf{x}^*, \mathbf{x}_1), \dots, C(\mathbf{x}^*, \mathbf{x}_l)]^T$.

2.2.3. Least-square support vector machine

Least-square support vector machine (LS-SVM) is a linear regression model: $y_i = \mathbf{w}^T \mathbf{x}_i + \epsilon_i$, where $\mathbf{w}$ is the model parameter, and ϵ_i is the regression error. As opposed to minimizing the sum of errors in conventional regression, LS-SVM includes a regularization term and aims to minimize the cost function of $Q = 1/2(\mathbf{w}^T \mathbf{w}) + \gamma \sum_{i=1}^l ||\epsilon_i||^2$, where the first part penalizes the magnitude of regression coefficients in order to avoid “over-fitting”, and the meta-parameter γ decides the relative weight regression error compared with the penalizing term (Thissen et al. 2004). The model parameters can be estimated by using the Lagrange method (Borin et al. 2006). The extension of above linear regression to non-linear models can be achieved by introducing the non-linear kernel function, for example the radial basis function (RBF).

3. Application study

3.1. Data sets

This comparative study is evaluated on three NIR datasets. The first was collected from the analysis of pharmaceutical tablets, which is publicly available at <http://www.models.life.ku.dk/research/data/Tablets/>. This dataset consists of 310 samples with 404 wavenumbers in the range of 7400-10500cm⁻¹. The objective is to determine the active substance content of the tablets. Two hundred samples were randomly selected as the training data for model development, and the remaining 110 samples were used for testing. Further details about this dataset are described in (Dyrby et al. 2002).

The second dataset is related to the transmittance spectra of wheat kernels (Pedersen et al. 2002). In this dataset, 523 samples from three different locations were analyzed by NIR spectroscopy at 100 wavelengths in the range of 850-1050nm. The data are available at http://www.models.life.ku.dk/research/data/wheat_kernels/. The objective of the analysis is to predict the protein content of the wheat kernels. This dataset is divided into 415 and 108 samples at random for training and testing, respectively.

The third dataset concerns about the gluten/starch powder mixture with 100 samples and 100 wavelengths (Martens, Nielsen, and Engelsen 2003). For each of five mixtures with different weight ratio of gluten and starch (1/0, 0.75/0.25, 0.5/0.5, 0.25/0.75, 0/1), five samples were loosely filled into different glass cuvettes and measured in two consecutive replicates. Then the samples were compressed firmly and measured in two consecutive replicates again. In this study, the training data consists of 60 random samples and the rest 40 samples are used for testing.

In order to give the reader a general idea of the datasets, one of the raw NIR spectra corresponding to the median of the response variable from the three datasets are shown in Fig. 1.

(Fig.1 about here)

3.2. Statistical evaluation of calibration methods

Nine chemometric calibration methods (D1, D2, SNV, EMSC, EISC, OPLEC, ANN, GP and LS-SVM) are considered in this comparative study. In addition to the individual methods, the combination of pre-processing and non-linear calibration techniques is also investigated. To keep this paper concise, we only consider two pre-processing techniques, SNV and EMSC, which are followed by developing non-linear models to give six combinations: SNV+ANN, EMSC+ANN, SNV+GP, EMSC+GP, SNV+LS-SVM and EMSC+LS-SVM. When dividing the limited dataset into training and testing data, the comparison results may be unreliable if the training and/or testing data are not representative of the whole dataset. The usual method to address this issue is to randomly partition the dataset multiple times (e.g. 50 times), and then apply each calibration method to obtain 50 RMSEPs. The average RMSEP from each method can be used for comparison purpose. Furthermore, a rigorous statistical hypothesis testing procedure, the paired t -test, is adopted to determine whether one method is statistically significantly better than the other (Couliden 1956). After calculating the t statistic based on the RMSEPs of two calibration methods, the corresponding p -value can be obtained. If $p < 0.05$, then we may conclude that these two calibration methods have attained significantly different RMSEPs.

3.3. Software

All computation was carried out in Matlab. The Matlab PLS toolbox version 4.2 (Eigenvector Research, Inc., Wenatchee, WA, USA) was used to perform PLS, D1, D2 and SNV. EMSC and EISC were performed by using the EMSC toolbox version 1.3 (Eigenvector Research, Inc., Wenatchee, WA, USA). The Matlab code for OPLEC is available in (Chen and Morris 2008). The Matlab Neural Network toolbox was employed for the implementation of ANN. LS-SVM was carried out using a toolbox from Katholieke Universiteit Leuven ([url: http://www.esat.kuleuven.ac.be/sista/lssvmlab/](http://www.esat.kuleuven.ac.be/sista/lssvmlab/)). The Matlab code for Gaussian process was described in (Rasmussen and Williams 2006) and is available at <http://www.gaussianprocess.org/gpml/code/matlab/doc/>.

4. Results and discussions

As mentioned in Section 2, pure spectra of individual chemical components are preferred as reference spectra for the implementation of EMSC (Chen, Morris, and Martin 2006). For the gluten/starch dataset, the 3rd and 93th samples are used as the pure spectra of gluten and starch respectively (Martens, Nielsen, and Engelsen 2003). However, for the other two datasets, no pure spectra are available, and thus the mean spectra of the entire training data are adopted as the reference spectra. In addition, whenever PLS is used in this study, either for calibration modelling or for estimating the OPLEC parameters, the number of the PLS components retained is always optimized through five-fold random-split cross-validation. Table 1 reports the number of selected PLS components, averaged over 50 random partitions of the dataset.

(Table 1 about here)

Prior to developing the non-linear calibration methods (ANN, GP and LS-SVM), PLS is first applied to reduce the dimensions of the data, which is a usual method to reduce the computational cost of parameter estimation in the non-linear models (Chen, Morris and Martin 2007; Gemperline, Long and Gregoriou 1991). The number

of retained PLS components is also reported in Table 1, where the dimension of data is reduced from several hundreds to less than 13, alleviating the computation for developing non-linear models. When carrying out LS-SVM, a Gaussian kernel function is used. The meta-parameters γ (see Section 2.2.3) and σ^2 (kernel parameter) are both optimized by five-fold random-split cross-validation.

Fig. 2 gives the RMSEPs averaged over 50 repeated random partitions of the three datasets for different calibration methods. The corresponding p -values of the paired t -test are also used to evaluate whether the results are statistically significant. For the tablets dataset, the general observation is that the non-linear calibration models attain lower prediction error than the pre-processing methods. The only exception is that ANN gives the same RMSEP of 0.31 as OPLEC and their paired p -value is 0.27 (>0.05), suggesting statistically insignificant difference between these two methods. Among the non-linear calibration methods, GP achieves the best predictive accuracy in terms of its RMSEP (0.27) and the p -values between GP and other methods being less than 0.001. LS-SVM also gives satisfactory RMSEP of 0.29. Among the pre-processing methods, OPLEC attains the best performance (RMSEP=0.31, $p<0.05$ when compared with other pre-processing methods); however, it does not outperform the non-linear models. The second best pre-processing method is EMSC (RMSEP=0.34) followed by D1, SNV and EISC; the latter three methods have the same RMSEP (0.37) and $p>0.05$ between them, indicating similar prediction performance. Finally, D2 gives the largest RMSEP of 0.44.

For the wheat kernel dataset, although RMSEPs of these techniques are similar, the pre-processing methods of EMSC, EISC, OPLEC and the non-linear methods of ANN, GP, LS-SVM still attain lower errors than D1, D2, SNV, and this better performances are statistically significant according to the paired t -test. Among these superior methods, GP and EMSC achieve the most accurate prediction. The p -value between GP and EMSC is greater than 0.05, indicating that they have the similar predictive capability. The RMSEP of 0.46 for EISC, OPLEC, ANN and 0.48 for LS-SVM suggest they also give satisfactory and similar predictions.

For the gluten/starch dataset, EMSC, OPLEC and GP give especially low prediction errors (RMSEP are 0.005, 0.003 and 0.005 respectively) and the p -values are all less than 0.001, implying that they achieve excellent predictive accuracy over other calibration methods. EISC also attains an acceptable performance in view of its RMSEP (0.010). The results of other calibration methods (D1, D2, SNV and ANN, LS-SVM) are unsatisfactory. It should be noted that the pure spectra of the chemical constituents are available for the application of EMSC and EISC for this dataset. If this is not the case in practice, their prediction accuracy is likely to decrease. Overall, the best pre-processing method (OPLEC) outperform the best non-linear model (GP) on this dataset.

(Fig. 2 about here)

To assess the stability of the calibration methods, the standard deviations of RMSEP are shown Fig. 3. For tablets dataset, the standard deviations of EISC and ANN are remarkably higher than the other methods, implying that EISC and ANN are sensitive to variations in the data. For wheat kernel dataset, EMSC, OPLEC,

ANN and GP appear to be more stable than the other techniques. For gluten/starch dataset, EMSC and OPLEC achieved outstanding stability in terms of very small standard deviation. Overall, no single method has shown consistently better stability than all other methods. The stability of regression models can be enhanced by combining the prediction from multiple models using, for example, the bootstrap aggregating (bagging) procedure (Chen and Ren, 2009; Tan, Qin, and Li. 2009). Further investigation on ensemble modelling for calibration is ongoing.

(Fig. 3 about here)

Fig. 4 shows the prediction performance when pre-processing and non-linear models are combined. For tablets dataset, SNV is preferred to EMSC as a pre-processing method to be combined with non-linear models. The combination of EMSC with the three non-linear calibration models gives significantly higher RMSEPs than the individual methods. For wheat kernel, only SNV+LS-SVM gives better results than the individual methods. Indeed, all the combinations attain similar RMSEPs. It may be concluded that for this data set, the use of SNV or EMSC as pre-processing for non-linear models does not introduce significant benefit. For gluten/starch dataset, although the combination of EMSC+ANN gives better result than ANN, it does not surpass EMSC. The RMSEPs of SNV+GP, SNV+LS-SVM and SNV+ANN are lower than SNV, but still higher than GP, LS-SVM and ANN. The predictive performances of other combinations are all worse than using the individual techniques.

(Fig. 4 about here)

Given the same amount of data, the computational time of the calibration methods mainly depends on the complexity of the model and the choice of cross-validation. Throughout this paper, the number of components in PLS was determined by using five-fold random-split cross-validation, with the range for validation being one to 15 components. Under these settings, D1, D2 and SNV are the most efficient in terms of computer time (all around 0.6 s). (All computational time was based on one partition of the tablet dataset as described previously. Computation was conducted on a Pentium 2.4 GHz computer running Windows XP). EMSC and EISC are slightly more complex (both around 1.5 s). OPLEC, despite its outstanding prediction performance, is the most time-consuming (127.3 s) because of the complicated strategy for parameter estimation. Non-linear ANN modelling requires to solve a non-linear optimization problem, resulting 1.9 s for its development. GP modelling takes around 4.8 s, mainly due to the need to invert a large matrix (details given in (Chen, Morris and Martin, 2007)). The very demanding computation of LS-SVM (90.1 s) is due to the need to cross-validate two continuous meta-parameters, which has been well recognized as a difficult task (Thissen et al., 2004). The computational cost of LS-SVM may be reduced by adopting alternative methods to find the optimal meta-parameters, such as Bayesian statistics (Van Gestel et al., 2002), a topic that is out of the scope of this paper.

5. Concluding remarks

This study compared the pre-processing and non-linear modelling techniques for the calibration of NIR spectroscopy in the presence of light scattering effect. A rigorous statistical procedure was adopted to obtain reliable comparison results. Although none of the techniques is always the best on all datasets, OPLEC and GP are found to be the most promising in terms of low prediction error. Compared with traditional approaches (D1, D2 and SNV), the more recently developed pre-processing methods (EMSC, EISC and OPLEC) are more favourable. This is due to better modelling of the light scattering effect (such as including the wavelength terms λ in the mode) and more advanced parameter estimation strategy (such as that of OPLEC). Therefore, if future research can lead to clearer understanding of the light scattering mechanism, and this information can be chemometrically modelled, then more powerful pre-processing techniques may emerge. However, in some practical situations, the light scattering effect is coupled with other disturbances to jointly affect the linearity of the spectral measurements. For example, when NIR spectrometers are applied for *in situ* monitoring of drying processes, both light scattering and temperature variation will affect the calibration accuracy. Under these circumstances, the development of pre-processing method by modelling of all major factors may be infeasible. In this regard, non-linear calibration techniques are preferred since they directly model the spectra-concentration relationship. Among the three non-linear models considered in this study, GP is recommended since it consistently attained lower RMSEP than ANN and LS-SVM. Finally, we have attempted to combine pre-processing and non-linear techniques; yet this strategy does not always outperform the individual techniques. This “hybrid” approach is conceptually appealing, and it will be further investigated in the future work.

Acknowledgment

The authors thank Professor Harald Martens at University of Copenhagen for supplying the gluten/starch powder mixture dataset.

References

- Balabin, R. M., R. Z. Safieva, and E. I. Lomakina. 2007. Comparison of linear and nonlinear calibration models based on near infrared (NIR) spectroscopy data for gasoline properties prediction. *Chemom. Intell. Lab. Syst.* 88(2): 183-188.
- Barnes, R. J., M. S. Dhanoa, and S. J. Lister. 1989. Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra. *Appl. Spectrosc.* 43(5): 772-777.
- Bishop, C.M. 1995. *Neural Networks for Pattern Recognition*. Oxford University Press.
- Borin, A., M. F. Ferrao, C. Mello, D. A. Maretto, and R. J. Poppi. 2006. Least-squares support vector machines and near infrared spectroscopy for quantification of common adulterants in powdered milk. *Anal. Chim. Acta.* 579(1): 25-32.
- Brown, C. D., L. Vega-Montoto, and P. D. Wentzell. 2000. Derivative preprocessing and optimal corrections for baseline drift in multivariate calibration. *Appl. Spectrosc.* 54(7): 1055-1068.
- Centner, V., J. Verdu-Andres, B. Walczak, D. Jouan-Rimbaud, F. Despagne, L. Pasti, R. Poppi, D. L. Massart, and O. E. De Noord. 2000. Comparison of multivariate calibration techniques applied to experimental NIR

- data sets. *Appl. Spectrosc.* 54(4): 608-623.
- Chen, T., J. Morris, and E. Martin. 2007. Gaussian process regression for multivariate spectroscopic calibration. *Chemom. Intell. Lab. Syst.* 87(1): 59-67.
- Chen, T., and J. Ren. 2009. Bagging for Gaussian process regression. *Neurocomputing* 72: 1605-1610.
- Chen, Z. P., and J. Morris. 2008. Improving the linearity of spectroscopic data subjected to fluctuations in external variables by the extended loading space standardization. *Analyst* 133(7): 914-922.
- Chen, Z. P., J. Morris, and E. Martin. 2006. Extracting chemical information from spectral data with multiplicative light scattering effects by optical path-length estimation and correction. *Anal. Chem.* 78(22): 7674-7681.
- Coulden, C. H. 1956. *Methods of Statistical Analysis*. 2nd eds. Wiley, New York.
- Dyrby, M., S. B. Engelsen, L. Norgaard, M. Bruhn, and L. Lundsberg-Nielsen. 2002. Chemometric quantitation of the active substance (containing C≡N) in a pharmaceutical tablet using near-infrared (NIR) transmittance and NIR FT-Raman spectra. *Appl. Spectrosc.* 56(5): 579-585.
- Geladi, P., D. Macdougall, and H. Martens. 1985. Linearization and scatter-correction for near-infrared reflectance spectra of meat. *Appl. Spectrosc.* 39(3): 491-500.
- Gemperline, P. J., J. R. Long, and V. G. Gregoriou. 1991. Nonlinear multivariate calibration using principal components regression and artificial neural networks. *Anal. Chem.* 63(20): 2313-2323.
- Gratton, E. 1971. Method for automatic correction of scattering in absorption spectra by using integrating sphere. *Biopolymers* 10(12): 2629-2634.
- Hadjiiski, L., P. Geladi, and P. Hopke. 1999. A comparison of modeling nonlinear systems with artificial neural networks and partial least squares. *Chemom. Intell. Lab. Syst.* 49(1): 91-103.
- Li, Y., J. Wang, Z.R. Chen, and X.T. Zhou. 2001. Artificial neural network for the quantitative analysis of air toxic VOCs. *Anal. Lett.* 34(12): 2203-2219.
- Long, J. R., V. G. Gregoriou, and P. J. Gemperline. 1990. Spectroscopic calibration and quantitation using artificial neural networks. *Anal. Chem.* 62(17): 1791-1797.
- MacKay, D. J. C. 1992. A practical Bayesian framework for back-propagation networks. *Neural Comput.* 4(3): 448-472.
- Martens, H., J. P. Nielsen, and S. B. Engelsen. 2003. Light scattering and light absorbance separated by extended multiplicative signal correction. Application to near-infrared transmission analysis of powder mixtures. *Anal. Chem.* 75(3): 394-404.
- Pedersen, D. K., H. Martens, J. P. Nielsen, and S. B. Engelsen. 2002. Near-infrared absorption and scattering separated by extended inverted signal correction (EISC): Analysis of near-infrared transmittance spectra of single wheat seeds. *Appl. Spectrosc.* 56(9): 1206-1214.
- Rasmussen, C. E., and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. MIT press: Cambridge, MA.
- Savitzky, A., and M. J. E. Golay. 1964. Smoothing and differentiation of data by simplified least squares procedures. *Anal. Chem.* 36(8): 1627-1639.
- Tan, C., X. Qin, and M. Li. 2009. Subspace regression ensemble method based on variable clustering for near-infrared spectroscopic calibration. *Anal. Lett.* 42: 1693-1710.

- Thennadil, S. N., H. Martens, and A. Kohler. 2006. Physics-based multiplicative scatter correction approaches for improving the performance of calibration models. *Appl. Spectrosc.* 60(3): 315-321.
- Thennadil, S. N., and E. B. Martin. 2005. Empirical preprocessing methods and their impact on NIR calibrations: a simulation study. *J. Chemom.* 19(2): 77-89.
- Thissen, U., B. Ustun, W. J. Melssen, and L. M. C. Buydens. 2004. Multivariate calibration with least-squares support vector machines. *Anal. Chem.* 76(11): 3099-3105.
- Thodberg, H. H. 1996. Review of Bayesian neural networks with an application to near infrared spectroscopy. *IEEE Trans. Neural Netw.* 7(1): 56-72.
- Van Gestel, T., J.A.K. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle. 2002. Bayesian framework for least-squares support vector machine classifiers, Gaussian processes, and kernel Fisher discriminant analysis. *Neural Comput.* 14: 1115-1147.
- Wang, Z. Y., J. N. Hwang, and B. R. Kowalski. 1995. ChemNets - theory and application. *Anal. Chem.* 67(9): 1497-1504.

Fig. 1. Overlay of the raw NIR spectra for the three datasets (solid line, tablets; dash-dot line, wheat kernels; dash line, gluten/starch powder mixture). Within each dataset, the spectrum corresponding to the median of the concentration value is selected for display. The arrows are directed to the corresponding axes.

Fig. 2. RMSEP of different calibration techniques for (a) tablets; (b) wheat kernel; (c) gluten/starch power mixture.

Fig. 3. Standard deviation of RMSEP for different calibration techniques: (a) tablets; (b) wheat kernel; (c) gluten/starch power mixture.

Fig. 4. RMSEP of combining pre-processing and non-linear calibration models for (a) tablets; (b) wheat kernel; (c) gluten/starch power mixture.

Table 1.

The number of selected components when using PLS in pre-processing or non-linear calibration techniques for (a) tablets; (b) wheat kernel; (c) gluten/starch power mixture. The values are averaged over 50 repeated random partitions of data.

(a)

Techniques	D1	D2	SNV	EMSC	EISC	OPLEC	ANN	GP	LS-SVM
Number	2.7	8.8	6.5	4.1	4.3	6.0	6.5	6.6	6.4

(b)

Techniques	D1	D2	SNV	EMSC	EISC	OPLEC	ANN	GP	LS-SVM
Number	11.1	9.3	12.1	9.4	9.2	12.0	12.0	12.1	12.0

(c)

Techniques	D1	D2	SNV	EMSC	EISC	OPLEC	ANN	GP	LS-SVM
Number	8.2	6.7	9.5	6.6	6.8	6.0	9.3	9.6	9.3