Preprints, 10th IFAC International Symposium on
Advanced Control of Chemical Processes
Shenyang, Liaoning, China, July 25-27, 2018

A spectroscopic chemometric modeling approach based on statistics pattern analysis

Devarshi Shah*, Q. Peter He*, Jin Wang*+

**Auburn University, Auburn, AL 36849 USA (e-mails:dshah@auburn.edu; qhe@auburn.edu; wang@auburn.edu).*

Abstract: Spectroscopic techniques such as near-infrared spectroscopy have gained wide applications in the last few decades. As a result, various soft sensors have been developed to predict sample properties from the sample's spectroscopic readings. Because the readings at different wavelengths are highly correlated, it has been shown that variable selection could significantly improve a soft sensor's prediction performance and reduce the model complexity. Currently, almost all variable selection methods focus on how to select the variables (*i.e.*, wavelengths or wavelength segments) that are strongly correlated with the dependent variable to improve the prediction performance. Although many successful applications have been reported, such variable selection methods do have their limitations, such as high sensitivity to the choice of training data, and poorer performance when testing on new samples. This is because the variables that are removed from model building may contain useful information about the sample property. To address this limitation, we propose a statistics pattern analysis (SPA) based method. Instead of selecting certain wavelengths or wavelength segments, the SPA-based method considers the whole spectrum which is divided into segments, and extracts different features over each spectrum segment to build the soft sensor. Two case studies are presented to demonstrate the performance of the SPA-based soft sensor and compared with a full partial least squares (PLS) model, and a synergy interval PLS (SiPLS) model.

Keywords: Soft sensor, Variable selection, Multivariate regression, Partial least squares, Statistics pattern analysis, NIR, UV/Vis, Chemometrics

1. INTRODUCTION

In the last few decades, spectroscopic techniques such as near-infrared (NIR) and UV/Vis spectroscopies have gained wide applications. Beyond their traditional applications in analytical chemistry, spectroscopic techniques are applied in many different fields, including biotechnological, pharmaceutical, petrochemical, and agricultural and food industries (Gendrin, 2008; Karoui and De Baerdemaeker, 2007; Meher et al., 2006). This is mainly due to their advantages over other analytical techniques, such as non-invasive and limited pre-treatment requirement. In order to correlate the spectroscopic readings of a sample with its properties of interest, multivariate regression methods such as multiple linear regression (MLR), principal component regression (PCR) and partial least squares (PLS) are commonly used to build data-driven models, often called soft sensors. These soft sensors allow the prediction of the interested properties of a sample based on its spectroscopic reading, such as estimating the moisture content in wheat using its NIR spectrum or estimating the concentration of individual components in a mixture using its UV/Vis spectrum.

Because the spectroscopic readings at different wavelengths are highly correlated, soft sensor development based on spectroscopic measurements is nontrivial. In fact, highly correlated regressor variables is the challenge to most soft sensor applications. Although multivariate regression methods based on dimension reduction approaches such as PCR and PLS have inherent capability of handling large number of correlated variables, it has been shown that variable selection, when combined with multivariate regression, can significantly improve the soft sensor's prediction performance, reduce the model complexity, as well as provide better insight into the nature of the process/system of interest (Wang et al., 2015) .

The spectroscopic measurements at different wavelengths contain a lot of information about the sample, which is why soft sensors can be developed to relate the spectrum to the sample properties of interest. Obviously, measurements at some wavelengths are highly correlated with the sample properties while the others are not. In addition, different wavelengths could contain different level of noises. Therefore, the goal of variable selection for spectroscopic data is to identify the subset of wavelengths that show the highest correlations to the interested properties of a sample to produce better estimate for new samples. Another potential benefit of variable selection is to eliminate measurements at wavelengths containing significant noises for better accuracy and performance of the soft sensor models (Xiaobo et al., 2010) .

Due to the benefits mentioned above, variable selection is viewed as a critical step in spectroscopic chemometrics model development and has drawn significant interest in the last few years. For example, Xiaobo et al. (2010) provided an excellent review of different variable selection methods for soft sensors using NIR data, and (Balabin and Smirnov, 2011) compared 17 different variable selection methods using a biodiesel dataset. Although variable selection, when done properly, often improves the model prediction performance, it does carry some limitations. As shown in our case studies presented in Section 3, variable selection can produce soft sensor models that are highly sensitive to the choice of training data, i.e., data used for model calibration. Due to the noises and unknown disturbances contained in the training data, the wavelengths selected to optimize the prediction performance based on the

training and validation data may be "tilted" to overfit or capture the noise or unknown disturbances contained in the training and validation data. As a result, the model prediction performance may deteriorate significantly when model is extrapolated or applied to new samples. In fact, this limitation is not unique to spectroscopic chemometrics models, it is true to all data-driven soft sensor models, which is in essence a balance between model accuracy and robustness. To help address this limitation, we propose a new soft sensor approach by adapting the statistics pattern analysis (SPA) framework we developed for process monitoring. In the SPA-based soft sensor modelling approach, the whole sample spectrum is divided into segments, and different features of each spectrum segment, instead of the spectrum readings themselves, are utilized to build the soft sensor model. In this way, the information contained in the whole spectrum is utilized while the effect of noise is removed or reduced. In addition, the number of variables used for model building is significantly reduced.

To demonstrate how the SPA based soft sensor approach could improve the model prediction accuracy and robustness, we compare its performance with synergy interval PLS (SiPLS), one of the most commonly used interval selection based methods (Norgaard et al., 2000; Silva and Wiebeck, 2017; Wang et al., 2012). The rest of the paper is organized as the following: Section 2 briefly reviews the SiPLS method and introduces the proposed SPA based soft sensor; Section 3 presents two cases studies; Section 4 draws the conclusion, and discusses the limitations of the proposed approach.

2. OVERVIEW OF INTERVAL SELECTION METHODS

Interval selection methods are variable selection approaches that explicitly or implicitly define intervals of the spectrum data in order to maintain a continuous variable selection. In this work, we focus on interval selection methods over other variable selection methods for comparison, not only because of their better performance in the cases studies (Norgaard et al., 2000; Silva and Wiebeck, 2017) , but also due to their fundamental roots in molecular chemistry. It has been well recognized that the general features of molecular spectra are of continuous bands rather than discrete responses. Therefore, it is reasonable to expect that a variable selection algorithm operating on a molecular spectrum would select regions of the spectrum rather than discrete wavelengths. In addition, the measured sample spectra are usually not aligned, so choosing an interval, instead of individual wavelength, would provide better containment of relevant information.

As an example, Fig. 1 plots the NIR spectra of a pharmaceutical tablet dataset (David W. Hopkins, 2003) . It can be seen that there are many clear absorption bands of the active pharmaceutical ingredient (API) from 600 to 1800 nm. It also shows that spectroscopic data usually contain large number of highly correlated spectral variables - because the general features of molecular spectra are of continuous bands, the neighbouring wavelengths of an absorption band are highly correlated to each other. Therefore, the wavelengths that are adjacent to each other offer similar information. This is why variable selection is highly desired for spectroscopic data. However, as shown in Fig. 1, the shape of different molecular spectra, i.e., the peaks corresponding to different molecular structures are different, which suggests that if only the peaks were chosen for model building, it may not capture sufficient information about the underlying molecular structure for accurate prediction. Last but not least, because noise and baseline drift are usually present in spectroscopic data, using an interval of wavelengths for model building could offer more robust prediction performance, without requiring extensive sample pre-processing.

2.1 Interval PLS and synergy interval PLS

Interval PLS (iPLS) method (Norgaard et al., 2000) is the most straightforward example of interval selection method, where the whole spectrum is divided into non-overlapping sections (intervals), as shown in Fig. 1. Then a separate PLS model is developed for each section and the PLS model that offers the best prediction performance will determine the most informative wavelength range. The interval width is the major tuning parameter of iPLS model, together with the order of each PLS model. Synergy interval PLS (SiPLS) (Norgaard et al., 2000) is an improved version of iPLS. Compared to iPLS where only a single interval is used for model building, SiPLS allows the combination of multiple intervals (2, 3 or 4) to be selected for model building. The tuning parameters for SiPLS include the width of the interval, the number of intervals to be combined and number of principal components (PCs) to retain. Because SiPLS is an improved version of iPLS, and provides improved performance over iPLS, in this work we only compare the performance of SiPLS with full PLS model and SPA-based method. The SiPLS algorithm used in this work was downloaded from www.models.life.ku.dk/iToolbox.

2.2 SPA-based soft sensor

It is a common belief that the spectrum peaks carry the most important information about the sample. However, even for the spectrum segments that do not contain obvious absorption peaks such as 750nm-1050nm in Fig. 1, they could contain important information about the sample. In order to retain as much information as possible from the sample spectra while significantly reduce the number of variables, as well as to remove/reduce the effect of measurement noise and based line drift, we propose the SPA-based soft sensor approach.

SPA is a process monitoring framework that the authors developed previously (He and Wang, 2010; Wang and He, 2010) , where the statistics of process variables, instead of the process variables themselves, are monitored to determine the process operation status. SPA offers many advantages such as effectively addressing process nonlinearity and non-Gaussianity, non-synchronized batch trajectories, etc. Its effectiveness and performance have been demonstrated in multiple case studies (He and Wang, 2010; Wang and He, 2010).

In this work, we adapt SPA to help address the challenge in variable selection for soft sensor model development. As

shown in Fig. 1, in the SPA-based approach we first divide the whole spectrum into non-overlapping intervals, which is

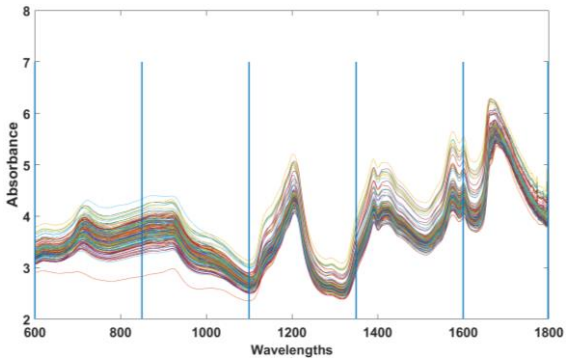

Figure 1. NIR spectra of pharmaceutical tablets

similar to SiPLS; then different features of each spectrum interval, such as the mean, standard deviation, skewness, kurtosis, are used as regressors to build the soft sensor model. In this way, information from the whole spectrum will be utilized for model building, but with significantly reduced number of variables. The schematic diagram of the proposed SPA-based soft sensor approach is shown in Fig. 2.

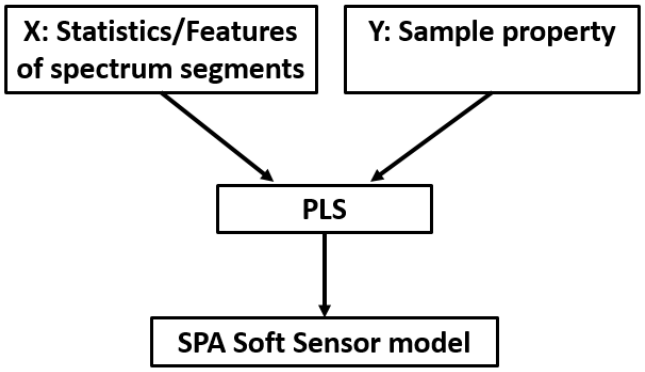

Fig. 2. Schematic of SPA based soft sensor model

There are several benefits associated with the SPA soft sensor. First, it utilizes the information from the whole spectrum to build the soft sensor model, which provides better model robustness; second, by extracting features of the spectrum segment in each interval, which involves computing the average of certain functions of absorption at different wavelength, it can reduce the effect of noise and baseline drift; third, the number of features is significantly smaller than the number of readings of the spectrum, which makes model development, optimization and update much faster compared to SiPLS.

3. CASE STUDIES

In this work, we use two case studies to examine the accuracy and robustness of the SPA-based soft sensors, and compare them with the full PLS model that utilize the whole spectrum, and SiPLS model. In the first case study, the soft sensor is developed to predict the individual cell concentration in a mixed culture using its UV/Vis spectrum. In the second case study, the soft sensor is developed to predict the API concentration in a pharmaceutical tablet.

To provide a fair comparison, we optimized all modelling approaches, and conducted Monte Carlo simulations whenever applicable. For each case study, the entire dataset is divided into three subsets: calibration set used to build the model; validation set used to optimize the model performance by tuning model parameters; testing set used to test the performance of the optimized model. The tuning parameters for each modelling approaches are the following: for the full PLS model, it is the number of PCs retained by the model; for the SiPLS model, they the number of intervals to be combined, the width of each interval and the number of PCs; for the SPA-based model, they are the width of each interval and the number of PCs. For the full PLS model and the SPA model, the number of PCs was determined by choosing PCs based on root mean squared error (RMSE) with $RSME_{n+1}/RSME_n > 0.99$ where n is the number of PCs. This criterion is essentially the same as the adjusted Wold's R criteria except that predicted residual error sum of squares (PRESS) is replaced with RMSE (Svante Wold, 1978) . For SiPLS, the first local minimum RMSE was used to determine the number of PCs, which is the default in the downloaded algorithm. In this work, for the SPA model, we used the mean, standard deviation, skewness and kurtosis to build the model, without further selection. It should be noted that the features to be included in the SPA model plays an important role for SPA-based soft sensor (He and Wang, 2010; Wang and He, 2010) . However, as an initial attempt, in this work we did not optimize the features to be used.

In this work, we use two indices to evaluate the performance of the different soft sensors: the mean prediction error (MPE) and the root mean squared error (RMSE), with their definitions listed below.

Mean prediction error (MPE):

$$MPE = \frac{\sum_{i=1}^{N}(Actual-Predicted)_i}{N} \quad (1)$$

Root mean squared error (RMSE):

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(Actual - Predicted)_i^2} \quad (2)$$

Where N is number of predicted values

3.1 Case study 1: Estimating individual biomass in mixed cultures

Due to many advantages associated with mixed cultures, the application of mixed cultures in biotechnology has expanded rapidly in recent years. However, how to efficiently and accurately monitor the individual cell populations in a mixed culture remains a challenging problem. The current approaches on individual cell mass quantification, such as cell counting or measuring ribosomal 16S DNA are time consuming and not suited for online monitoring. To address this difficulty, recently we developed a fast and accurate 'soft sensor' approach for estimating individual cell concentrations

In mixed cultures (Stone et al., 2017) . The developed approach utilizes optical density scanning spectrum (UV-Vis) of a mixed culture sample measured by a spectrophotometer to

estimate individual cell concentration through a PLS model. In Stone et al. (2017), the PLS soft sensor that utilized the whole spectrum without variable selection was shown to provide satisfactory performance, with significantly better precision and accuracy compared to cell counting method.

In this case study, one dataset reported in Stone et al. (2017) is used for evaluating the prediction performances of soft sensors based on full PLS model, SiPLS model and SPA-based model. The dataset consists of optical density spectra of 47 samples of mixed *E. coli* and *S. cerevisiae* cultures with known individual cell mass composition. The UV/Vis spectra of all 47 samples are plotted in Fig. 3(a), which clearly shows that the 47 samples can be divided into 6 groups. Such grouping is due to the experimental design. Within each group, the cell concentration of each individual strain changes in the opposite direction linearly, that is, the linearly increasing concentration of *E. coli* is paired with linearly decreasing concentrations of *S. cerevisiae*, while the total OD at 670nm is maintained at a fixed level for each group. Details on the experimental design and sample preparation can be found in Stone et al. (2017). When we apply PCA to analyse the dataset, such grouping can be clearly seen in the score plot, as shown in Fig. 3 (b), where the sample spectra are projected onto the 2-dimentional principal component subspace. To examine the robustness of different soft sensor models, we consider two different scenarios, denoted by A and B, where samples from the same group of the testing samples are included or not included in the model calibration respectively.

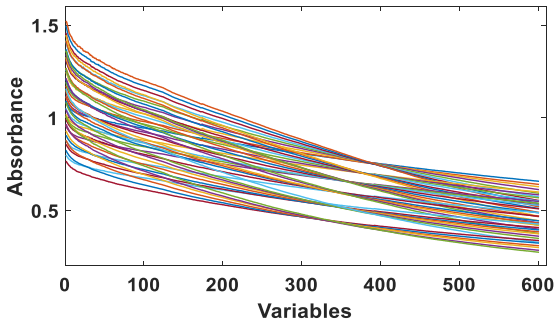

Fig 3(a): UV/Vis spectra of co-culture

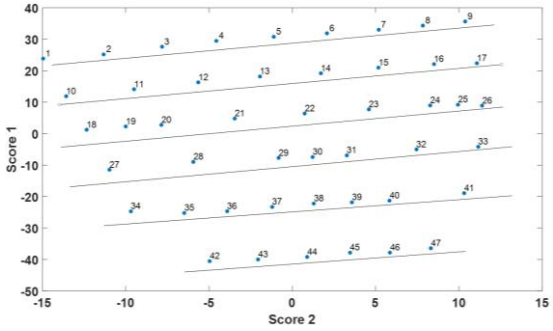

Fig 3(b): Score plot of spectra.

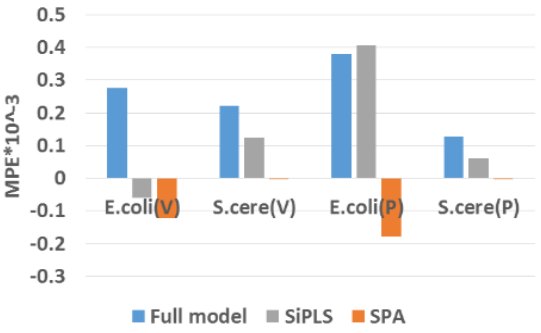

Fig. 4(a). MPE comparison for scenario A. (V= Validation results, P=Prediction results)

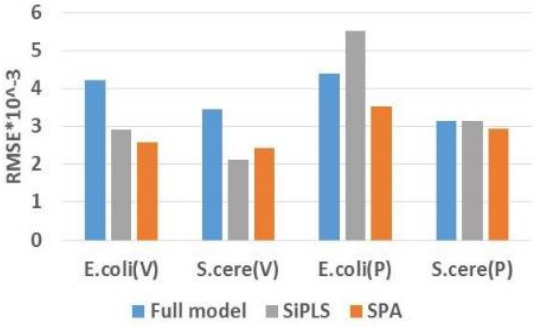

Fig. 4(b). RMSE comparison for scenario A. (V= Validation results, P=Prediction results)

3.1.1 Scenario A

In this scenario, the whole dataset was randomly divide into calibration, validation and test subsets, with each containing 20, 12 and 15 samples respectively. In addition, at least 2 samples from each group were randomly selected to be included in the calibration subset. This is to ensure that the training dataset captures all the groups (variations) of the test data. To examine the robustness of each modelling approach with respect to the selection of calibration dataset, 100 Monte Carlo runs were conducted, and totally 1500 predictions (15 samples by 100 runs) are pooled together to compute the performance indices. Fig. 4 compares the performance of the three soft sensors, where Fig. 4 (a) compares MPE while Fig. 4(b) compares RMSE from three approaches. For this scenario, all three models offer satisfactory performance, with fairly small MPE and RMSE. Fig. 4 also shows that although both SiPLS and SPA perform better than full PLS on the validation set for both strain concentrations, only SPA shows improved performance on the testing set, while SiPLS shows worse performance on estimating the *E. coli* concentration. Due to limited space, only the estimation results from SPA for the testing set are shown in Fig. 5, which demonstrate the unbiased estimate from SPA.

3.1.2 Scenario B

In this scenario, the first two groups (17 samples) were used as the test set. The remaining four groups were divided into calibration and validation sets containing 20 and 10 samples respectively. In addition, at least 2 samples from each of the four groups were randomly selected to be included in the calibration subset. 100 Monte Carlo runs were conducted and totally 1700 predictions (17 samples by 100) were pooled together to compute performance indices. Fig. 6 compares the performance of the three soft sensor models for scenario B, which shows that when the testing data are from different groups than the training and validation data, model prediction performance deteriorates significantly for all three models. However, SPA shows the least deterioration and offers essentially unbiased estimates. In contrast, the full PLS model

and the SiPLS model show significant bias in estimating *E. coli* concentrations. This is confirmed by Fig. 7, which plots the estimated *E. coli* concentration from the three models.

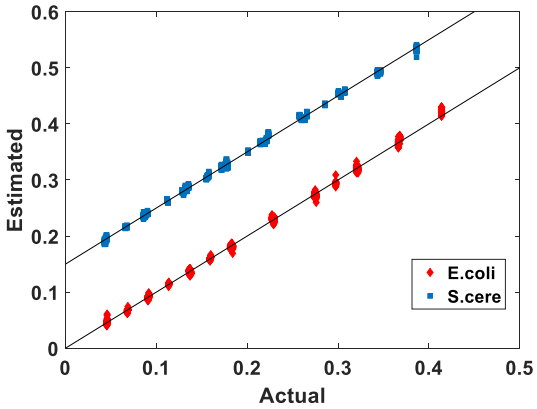

Fig. 5. SPA estimates for scenario A

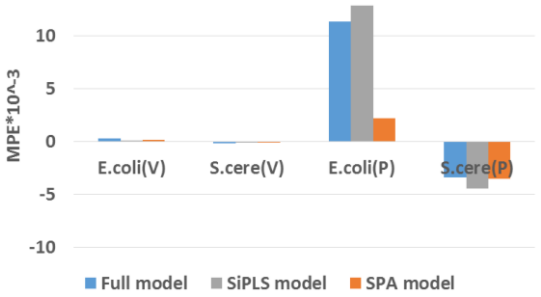

Fig. 6(a). MPE comparison for scenario B. (V= Validation results, P=Prediction results)

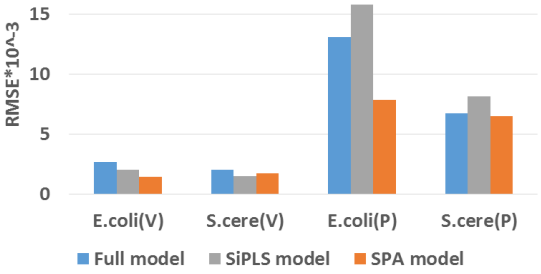

Fig. 6(b). RMSE comparison for scenario B. (V= Validation results, P=Prediction results)

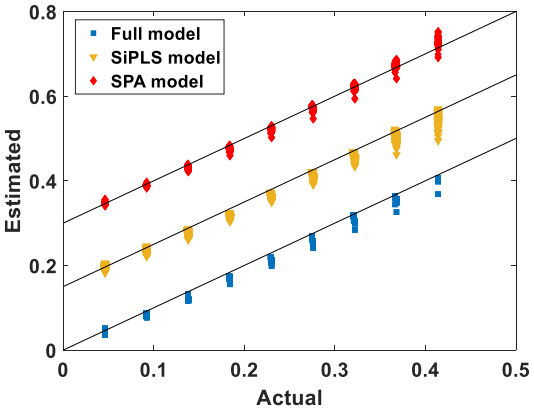

Fig. 7 Estimate of *E. coli* concentration for scenario B (estimated values of SiPLS and SPA shifted for clarity)

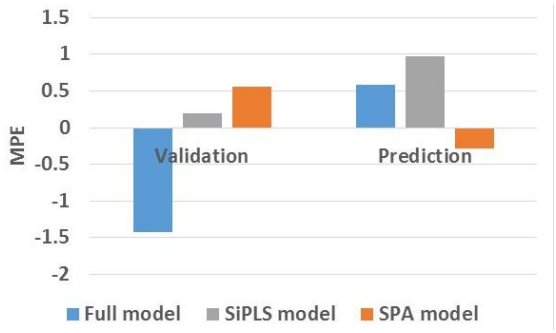

Fig. 8(a). Mean Error comparison for case study-II.

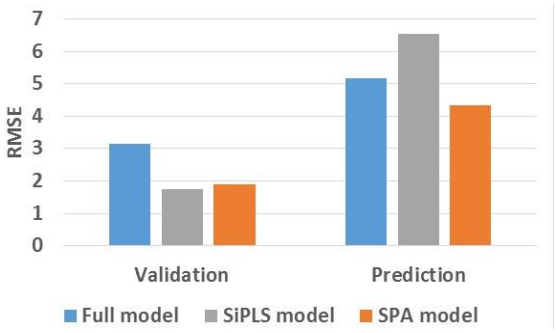

Fig. 8(b) RMSE comparison for case study-II

Table 1. Performance improvement by SiPLS and SPA

	model	MPE	% imp	RMSE	% imp
Valid.	**Full**	-1.42	0.0	3.13	0.0
	SiPLS	0.20	86.1	1.74	44.6
	SPA	0.56	61.0	1.90	39.3
Pred.	**Full**	0.59	0.0	5.18	0.0
	SiPLS	0.97	-64.6	6.55	-26.3
	SPA	-0.28	52.9	4.34	16.2

3.2 Case study II: API estimation using NIR of Pharmaceutical tablet

The pharmaceutical dataset, which was downloaded from http://www.eigenvector.com/data/tablets/index.html, consists of NIR spectra of 654 pharmaceutical tablets, and was divided into three subsets: 154 samples in calibration, 40 samples in validation and 460 samples in testing.

Fig. 8 compares the three modelling approaches on the pharmaceutical data sent, for both validation and prediction subsets. Again, although both SiPLS and SPA model showed improved performance over the full model on the validation subset, only SPA showed significantly improved performance on the testing subset. Table 1 lists the percentage improvement (if positive) or deterioration (if negative) with the full model as the reference. Table I shows that SPA model delivers 52.9% reduction in MPE and 16.2% reduction in RMSE, while SiPLS shows 64.6% and 26.3% deterioration, respectively. Fig. 9 plots the predictions from the three models and compared with the measurements. The two vertical ovals highlights the segments where SPA predictions are significantly better than

the other two approaches.

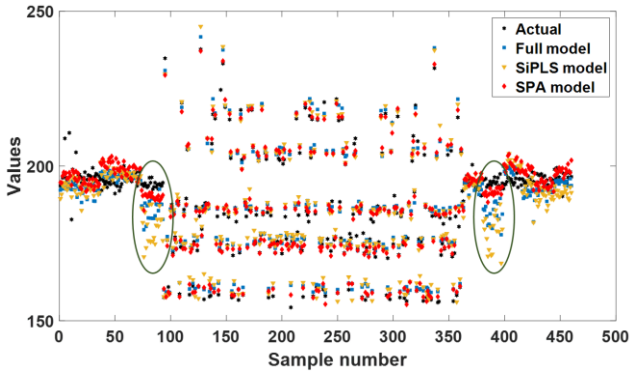

Fig. 9 Comparison of estimates and actual measurements of all test samples.

4. CONCLUSIONS

Variable selection has been recognized as a critical step in soft sensor development, particularly for chemometrics models. Up to date many variable selection methods have been developed to improve soft sensor prediction performance.

In this work, we present an SPA-based chemometrics soft sensor. Instead of selecting certain wavelengths or wavelength segments, the SPA-based method considers the whole spectrum which is divided into segments, and choose different features over each spectrum segment to build the soft sensor. In this way, it can not only significantly reduce the number of independent variables, but also utilize the information contained in the whole spectrum while reducing noises. Two case studies demonstrate the performance of the SPA based soft sensor approach. When the model is extrapolated to test samples that are different from training data which is common in really applications, the SPA based approach demonstrates the most significant improvement over the full PLS model and the SiPLS model.

Since this is our initial effort in expanding the SPA framework to soft sensor development, we have not examined the contribution of each feature to the soft sensor models. It should be noted that the choice of the features to be included in the soft senor would play a key role in soft sensor performance and will be investigated in future studies.

REFERENCES

Balabin, R.M., Smirnov, S.V., 2011. Variable selection in near-infrared spectroscopy: Benchmarking of feature selection methods on biodiesel data. Anal. Chim. Acta 692, 63–72. doi:10.1016/j.aca.2011.03.006

David W. Hopkins, 2003. Shoot-out 2002: transfer of calibration for content of active in a pharmaceutical tablet. NIR news 14, 10–13.

Gendrin, C., 2008. Monitoring galenical process development by near infrared chemical imaging: One case study. Eur. J. Pharm. Biopharm. 68, 828–837. doi:10.1016/j.ejpb.2007.08.008

He, Q.P., Wang, J., 2010. Statistics pattern analysis: A new process monitoring framework and its application to semiconductor batch processes. AIChE J. 57, 107–121. doi:10.1002/aic.12247

Karoui, R., De Baerdemaeker, J., 2007. A review of the analytical methods coupled with chemometric tools for the determination of the quality and identity of dairy products. Food Chem. 102, 621–640.

Meher, L., Vidyasagar, D., Naik, S., 2006. Technical aspects of biodiesel production by transesterification—a review. Renew. Sustain. Energy Rev. 10, 248–268. doi:10.1016/j.rser.2004.09.002

Norgaard, L., Saudland, A., Wagner, J., Nielsen, J.P., Munck, L., Engelsen, S.B., 2000. Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. Appl. Spectrosc. 54, 413–419.

Silva, D.J. da, Wiebeck, H., 2017. Using PLS, iPLS and siPLS linear regressions to determine the composition of LDPE/HDPE blends: A comparison between confocal Raman and ATR-FTIR spectroscopies. Vib. Spectrosc. 92, 259–266. doi:10.1016/j.vibspec.2017.08.009

Stone, K.A., Shah, D., Kim, M.H., Roberts, N.R.M., He, Q.P., Wang, J., 2017. A novel soft sensor approach for estimating individual biomass in mixed cultures. Biotechnol. Prog. 33, 347–354. doi:10.1002/btpr.2453

Svante Wold, 1978. Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models. Technometrics 20, 397–405.

Wang, J., He, Q.P., 2010. Multivariate Statistical Process Monitoring Based on Statistics Pattern Analysis. Ind. & Eng. Chem. Res. 49, 7858–7869. doi:10.1021/ie901911p

Wang, X., Bao, Y., Liu, G., Li, G., Lin, L., 2012. Study on the Best Analysis Spectral Section of NIR to Detect Alcohol Concentration Based on SiPLS. Procedia Eng. 29, 2285–2290. doi:10.1016/j.proeng.2012.01.302

Wang, Z.X., He, Q.P., Wang, J., 2015. Comparison of variable selection methods for PLS-based soft sensor modeling. J. Process. Control. 26, 56–72.

Xiaobo, Z., Jiewen, Z., Povey, M.J.W., Holmes, M., Hanpin, M., 2010. Variables selection methods in near-infrared spectroscopy. Anal. Chim. Acta 667, 14–32. doi:10.1016/j.aca.2010.03.048