A Feature-based Soft Sensor for Spectroscopic Data Analysis

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

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

Abstract

In the last few decades, spectroscopic techniques such as near-infrared (NIR) and UV/Vis spectroscopies have gained wide applications. As a result, various soft sensors have been developed to predict sample properties from its 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 deteriorated performance when testing on new samples. One possible reason is the removal of useful wavelengths or segments of wavelengths during the calibration process, which could be "tilted" to overfit or capture the noise or unknown disturbances contained in the calibration data. As a result, the model prediction performance may deteriorate significantly when the model is extrapolated or applied to new samples. To address this limitation, we propose a feature-based soft sensor approach utilizing statistics pattern analysis (SPA). 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. In other words, the SPA model contains the complete information from the full spectrum without any selection or removal, which we believe is the main reason for the high robustness of the SPA-based method. In addition, we propose a Monte Carlo validation and testing (MCVT) procedure and three MCVT-based performance indices for consistent and fair comparison of different soft sensor methods across different datasets. The MCVT procedure and indices are generally applicable for model comparison in other applications. Four case studies are presented to demonstrate the performance of the feature-based soft sensor and to compare it with a full partial least squares (PLS), a least absolute shrinkage and selection operator (Lasso), and a synergy interval PLS (SiPLS) based models following the proposed MCVT procedure. In addition, we examine the potential of kernel PLS (KPLS) based soft sensor approaches, examine their performances, and discuss their pros and cons.

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

1 Introduction

Spectroscopic techniques such as near-infrared (NIR) and UV/Vis spectroscopies have gained wide applications in the last few decades due to their advantages over other analytical techniques, such as non-invasiveness and low pre-treatment requirement. Beyond their traditional applications

in analytical chemistry, spectroscopic techniques have been applied in many different fields to determine properties such as octane number [1], moisture content [2], active chemicals in a samples [3], and microorganism concentration [4]. In order to correlate the spectroscopic readings of a sample to its properties of interest, multivariate regression models, also known as soft sensors, are often developed, which usually utilize multivariate statistical methods such as multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS) [5,6] and canonical variate analysis (CVA) [7]. Interestingly, although CVA identifies directions of maximum correlation between response and regressor variables while PLS may theoretically include directions that are irrelevant to response variable(s) [7,8], PLS is the most commonly used soft sensor platform and there seems no research showing the advantage of CVA over PLS for soft sensor modeling. In addition, when variable selection is implemented prior to soft sensor modeling, it is expected that the variable selection process would exclude regressor variables/features that are irrelevant to the response variable(s). Therefore, in this work we choose to use PLS as the modeling backbone for the proposed approach, although it is straightforward to extend the method to CVA based soft sensor. Meanwhile, nonlinear soft sensors that utilize artificial neural network (ANN) or kernel-based methods such as support vector regression (SVR), kernel-PLS (KPLS), etc. have also been proposed in the literature [9,10]. As most spectroscopic datasets have relatively small sample size (e.g., three out of four datasets used in this work have only 21-36 training samples and 16-28 validation samples), they are not sufficient to train a good NN model. Therefore, in this work, we examine KPLS based nonlinear soft sensor. It has been shown that KPLS is a very effective soft sensor approach competitive with other kernel-based approaches such as SVR [11–13]. Other advantages of KPLS include its robustness and straightforward generalization, and ease of tuning of the parameters [14].

For absorption spectroscopic measurements, absorbance values at different wavelengths correspond to light absorbed by different components of a sample as illustrated in Figure 1, where

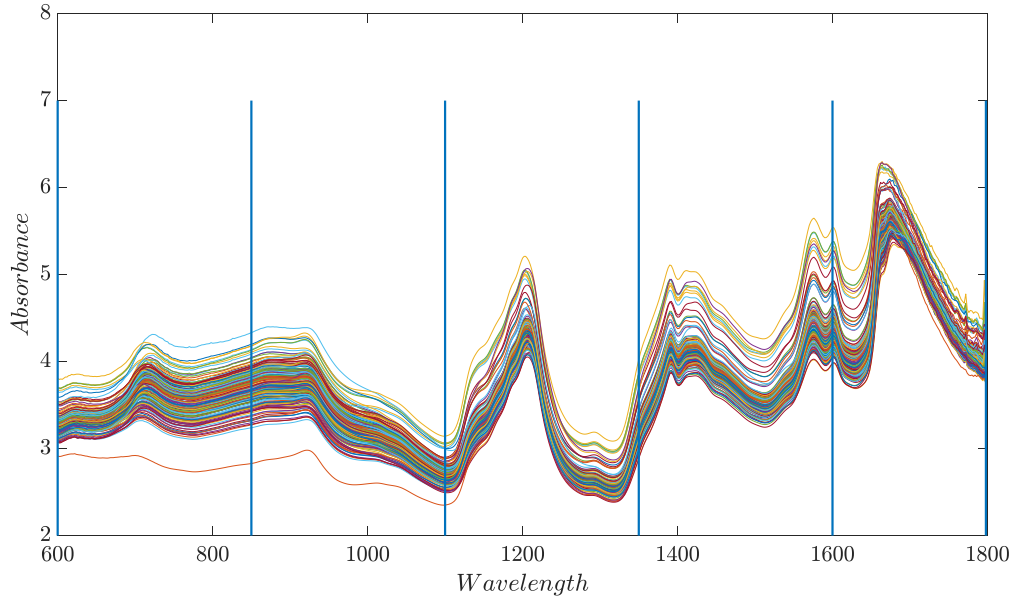


Figure 1. NIR spectra of pharmaceutical tablets (different colors refer to different samples)

the NIR spectra of a pharmaceutical tablet dataset are shown. It can be seen from Figure 1 that there are many clear absorption bands of the active pharmaceutical ingredient (API) from 600 to 1800 nm. Since the number of variables, which equals to the number of wavelengths where absorbances are measured, are usually large, substantial number of samples are required for building robust models. However, in some applications, the number of samples are limited. In those cases, the so-called "curse of dimensionality" would affect the predictive power of the model, where insufficient number of samples (compared to the number of variables) are used to build the model. On the other hand, it is well known that absorbance values at different wavelengths are not equally important in building such models. In addition, as shown in Figure 1, the absorbances of adjacent wavelengths offer similar information – because the general features of molecular spectra are of continuous bands. In other words, spectroscopic data contain large number of redundant or highly correlated spectral variables. 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 [10,15–17]. The goal of variable selection for spectroscopic data is to identify the subset of wavelengths that are closely related to the interested properties of a sample such that the model built using the subset of the wavelengths can better estimate the properties 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 [18].

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, Zou et al. (2010) provided a review of different variable selection methods for soft sensors using NIR data [18], and Balabin and Smirnov (2011) compared 16 different variable selection methods using a biodiesel dataset [16]. It is worth noting that there are many applications of variable selection in other areas [19–24], but it is not the focus of this work to review them here. Although variable selection, when done properly, often improves the spectrum model prediction performance, it does carry some limitations. As shown in the case studies presented in this work, variable selection can produce soft sensor models that are sensitive to the choice of training and validation 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 calibration performance based on the training and validation data may be "tilted" to overfit or to capture the noise or unknown disturbances contained in the calibration 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; instead, 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 feature-based soft sensor approach by adapting the statistics pattern analysis (SPA) framework we developed for process monitoring. In the SPA enabled feature-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 but the number of variables used for model building is

significantly reduced. In addition, the effect of noise can be reduced or removed in the feature extraction process. The performance of the proposed method is extensively tested in this work and is compared with a full PLS model utilizing all variables, a shrinkage method least absolute shrinkage and selection operator (Lasso) and an interval based variable selection method synergy interval PLS (SiPLS) [25]. In addition, we explore nonlinear KPLS based soft sensor applied to the original full spectra, as well as to the SPA features. We examine their performances and discuss their pros and cons. Four datasets from different fields, including agriculture, petroleum, pharmaceutical and biochemical, are chosen to show the versatile applicability of the proposed feature-based approach. For consistent and fair comparison, we propose a Monte Carlo validation and testing (MCVT) procedure and three MCVT-based performance indices for evaluating the performance of different soft sensor methods across different datasets. It is worth noting that the MCVT procedure and the MCVT-based performance indices are generally applicable for model comparison in other applications.

The rest of the paper is organized as follows. Section 2 reviews PLS and KPLS based soft sensors applied to spectroscopic data. Section 3 reviews Lasso and SiPLS based variable selection methods. Section 4 presents the proposed feature-based soft sensor. The datasets used in this study are introduced in Section 5. Section 6 presents the proposed Monte Carlo validation and testing (MCVT) procedure and three MCVT based soft sensor performance indices. Section 7 discusses results of the case studies and Section 8 draws conclusions.

2 Brief review of PLS and KPLS based soft sensors

As discussed in Sec. 1, there are many linear and nonlinear soft sensors developed in the literature for spectroscopic data analysis. Here we briefly review PLS and KPLS based soft sensors studied in this work.

2.1 PLS based soft sensor

The construction of PLS based soft sensor using full or selected wavelengths follows the same procedure as follows. First, the spectra are used to construct the independent variable matrix $\boldsymbol{X}$, which has the dimension of $n \times p$, where n is the numbers of samples, and p is the number of all or selected wavelengths. Each row of $\boldsymbol{X}$ corresponds to a sample spectrum of absorbance at all or selected wavelengths. Then, the physical or chemical properties of the samples are used to construct the dependent variable vector y with the dimension of $n \times 1$ where the number of properties is 1. Finally, a linear PLS regression model is developed to correlate y with $\boldsymbol{X}$ by $y = \boldsymbol{X}\beta + \varepsilon$, where β is a $p \times 1$ regression coefficient vector, and ε is a noise term for the model which has the same dimensions as y. Usually, the variables in $\boldsymbol{X}$ and y are centered by subtracting their means and scaled by dividing by their standard deviations. The formulation can be straightforwardly extended to the cases with multiple properties. In this work, the *plsregress* function from Matlab Statistics and Machine Learning Toolbox is used to build the PLS models.

2.2 KPLS based soft sensor

Many variations of nonlinear PLS have been proposed in literature [26–29], among which KPLS has been extensively studied [11,12,30]. In KPLS, the independent variable matrix $\boldsymbol{X}$ is mapped onto a higher dimensional feature space using a non-linear function $\boldsymbol{\Phi}(\boldsymbol{X})$. This mapped data

appears as a dot products of the mapping functions in dual space, and thus by applying "kernel trick" the dot product of the mapping function can be replaced by kernel function ($\boldsymbol{K} = \boldsymbol{\Phi}\boldsymbol{\Phi}^T$). A PLS soft sensor is then constructed in the mapped space corresponding to a nonlinear function in the original input space. More detailed derivations and discussions of KPLS can be found in [11,12,31,32]. In this work a radial basis function (RBF) or Gaussian kernel shown below is deployed.

$$\boldsymbol{K}(x_i, x_j) = exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (1)$$

where σ is the kernel parameter.

3 Brief review of Lasso and SiPLS based variable selection methods

As discussed in Sec. 1, there are many variable selection methods developed in the literature for spectroscopic data analysis. In this section we briefly review Lasso and SiPLS based variable selection methods studied in this work.

3.1 Least absolute shrinkage and selection operator (Lasso)

Lasso selects variables by minimizing the square of the L^2 norm of the residual vector with a penalty on the L^1 norm of the coefficient vector [33] as below.

$$J(\beta) = \underset{\beta}{\operatorname{argmin}}(\|y - \boldsymbol{X}\beta\|_2^2 + \lambda\|\beta\|_1)$$

where λ is a nonnegative regularization parameter to be optimized during model calibration or cross-validation. More detailed discuss on Lasso algorithm can be found in [33], and comparison of Lasso to other variable selection methods for PLS-based soft sensors in [34]. In this work, the *lasso* function from Matlab Statistics and Machine Learning Toolbox is used.

3.2 Interval based variable selection approaches

One of the early work under this category is called interval partial least squares (iPLS) [25]. In this method, a whole spectrum is divided into *N* non-overlapping segments of the same size (except the last segment) as shown in Figure 1. For each segment, a PLS model is developed. In order to find the segment that provide the best performance, the procedure is repeated with different value of *N*'s, and a standard cross-validation approach such as RMSECV (root mean squared error of cross-validation) can be used for performance evaluation. The segment with the best performance on the validation data, together with the optimal parameters such as the number of principal components (PCs), are used to develop a final PLS model for prediction on the test data. Case studies have shown that the performance of iPLS is comparable, rather than outperforming, other variable selection based methods such as principal variable (PV), forward stepwise selection (FSS) and recursively weighted regression (RWR). Since then, several variations of iPLS have been proposed, including backward interval PLS (biPLS), Moving window PLS (mwPLS), and Synergy interval PLS (SiPLS) [25,35]. In biPLS, the spectrum is divided into *N* segments and a PLS model is developed by leaving out one interval at a time. Intervals are removed until last, best performing interval is identified. In mwPLS, PLS models are calculated based on moving window approach, size of the window is fixed and window with best performing model is considered for future

predictions. SiPLS is another improved version of iPLS [25]. Compared to iPLS where only a single interval is used for model building, SiPLS allows the combination of multiple segments (*e.g.*, 2, 3 or 4) to be selected for model building. The tuning parameters for SiPLS include the number of segments the spectrum to be divided into (*i.e.*, *N*), the segments to be included in the model, and the number of principal components (PCs) to be retained in the model. In this study, only results obtained using SiPLS is compared as it was shown that SiPLS outperforms other interval based approaches [25]. The SiPLS Matlab code used in this work was downloaded from www.models.life.ku.dk/iToolbox.

4 The proposed Statistics pattern analysis (SPA) enabled feature-based soft sensor

As discussed above, good variable selection methods result in reduced models that are simpler, more robust and provide better prediction performance. However, when the training samples are not properly selected or there are not sufficient samples to cover the entire range of the properties to be predicted, the variable selection may be biased towards the covered property region while the extrapolability of the model can be poor. This can be evaluated by comparing the performance of the model on the validation samples to that of the test samples. A significant deterioration of performance on the test samples compared to that of the validation samples would indicate such a deficiency in variable selection. To address this potential issue while still significantly reducing the number of variables, we propose a feature-based soft sensor using statistics pattern analysis (SPA).

Statistics pattern analysis (SPA) is a process monitoring framework that the authors developed previously [36–38], in which 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 in process monitoring have been demonstrated in multiple case studies [36–38].

In the original SPA based process monitoring approach, the statistics are calculated along the time dimension and PCA is performed on the statistics for fault detection and diagnosis. There is no response variable involved. Also, the statistics cannot be obtained on an individual sample. There must be a group/window of samples in order to estimate the statistics. In the proposed SPA feature-based soft sensor, the statistics are calculated along the variable (i.e., wavelength) dimension and the statistics are correlated to response variable(s) through PLS. In this case, statistics is estimated based on an individual sample and the properties are estimated individually for each test sample. Specifically, as shown in Figure 2, in the SPA-based soft sensor we first divide each spectrum into s non-overlapping segments, which is similar to SiPLS; then f different features are extracted from each spectrum segment for each sample, which are raw spectrum readings without any scaling. The extracted features, such as the mean, standard deviation, skewness, kurtosis, are used as the regressors (totally $s \times f$ features for each sample) to build the soft sensor model. With n samples, the dimension of X would be $n \times (s \times f)$ and the dimension of Y would be $n \times 1$ for a single property, or $n \times m$ for m properties. Both X and Y are auto-scaled to zero mean and unit variance for PLS modeling. The spectrum segmentation intervals (or number of segments),

statistics used for model building, and number of PC's for PLS are optimized based on cross validation. 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 feature-based soft sensor approach is shown in Figure 2.

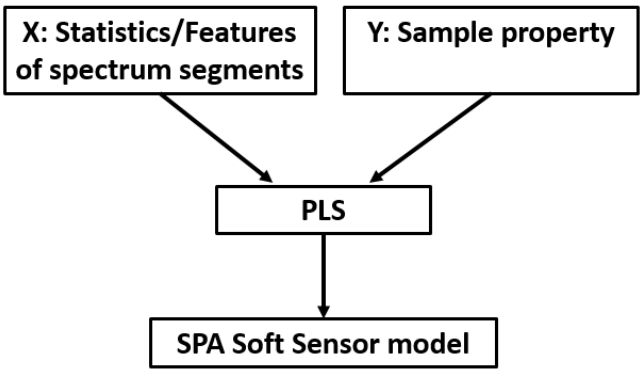


Figure 2. Schematic of SPA feature-based soft sensor

There are several benefits associated with the SPA feature-based 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 reduces the effect of noise; finally, it offers the flexibility to utilize nonlinear features and higher order statistics to better capture the nonlinear relationships between sample absorbance spectrum and property, although it is worth noting that there are other ways to capture the data nonlinearity (e.g., [39]). In addition, a nonlinear regression method such as KPLS can be used in place of PLS to further capture the nonlinear relationships, if any, between SPA features and properties of interest.

In this study the following eight different features/statistics are considered as the candidate features to be modeled: mean (μ), standard deviation (σ) or Variance (σ^2), skewness (γ), kurtosis (κ), average of first derivative of spectrum over an interval (AFD), average of second derivative of spectrum over an interval (ASD), slope of linear regression line (SLL) and coefficient of squared term for second order regression line (SSL). μ and σ represent the overall change in a given spectrum segment. γ, κ, SLL and SSL provide information on different aspects that characterize the shape of the spectrum in an interval. AFD and ASD give rate of change and rate of rate of change of absorbance spectrum with respect to wavelength. Note that the first and second derivatives of the absorbance spectrum, instead of spectrum itself, have been used for spectral analysis [40–42].

5 Case studies

In order to comprehensively compare SPA with PLS full model and reduced models based on Lasso and SiPLS, four datasets from different fields are used in this work.

1. **Corn dataset:** This dataset consists of 80 samples of NIR absorbance spectra from three spectrometers and corresponding property values of moisture, oil, protein and starch. Wavelength range is 1100nm to 2498nm at 2nm interval. In this paper moisture property and

NIR spectra from mp6spec was used for study and comparison, any other property can also be used. More detailed discussion of the dataset can be found in [43].

2. **Gasoline dataset:** This dataset consists of 60 samples of NIR absorbance spectra and corresponding octane number. Wavelength range is 900nm to 1700nm at 2nm interval. More detailed discussion of the dataset can be found in [44].
3. **Pharmaceutical tablets dataset:** This dataset consists of 655 samples of NIR absorbance spectra of pharmaceutical tablets and corresponding values of total weight, hardness and Active pharmaceutical ingredients (API). Wavelength range is 600nm to 1798 nm at 2nm interval. In this paper API of the tablets was used for comparison and study. This dataset has already been divided into calibration, validation and test sets. More detailed discussion of the dataset can be found in [45,46].
4. **Co-culture dataset:** This dataset consists of 47 samples of UV/Vis absorbance spectra of *E.coli* and *S. cerevisiae* co-culture with known individual cell mass concentration. In this data spectra were clearly separated into 6 groups. Wavelength range is 300nm to 900nm at 1nm interval. Detailed description of the dataset and the experimental design can be found in [4].

Spectra of all four datasets used for this study are shown in Figure 3.

For consistent and fair comparison across different datasets, the datasets were divided into training, validation and test sub-sets in consistent proportions. For small datasets (i.e., corn, gasoline and coculture datasets), about 20% of all samples were left out as test samples while the remaining ~80% of all samples were used for model calibration (i.e., training and validation). For

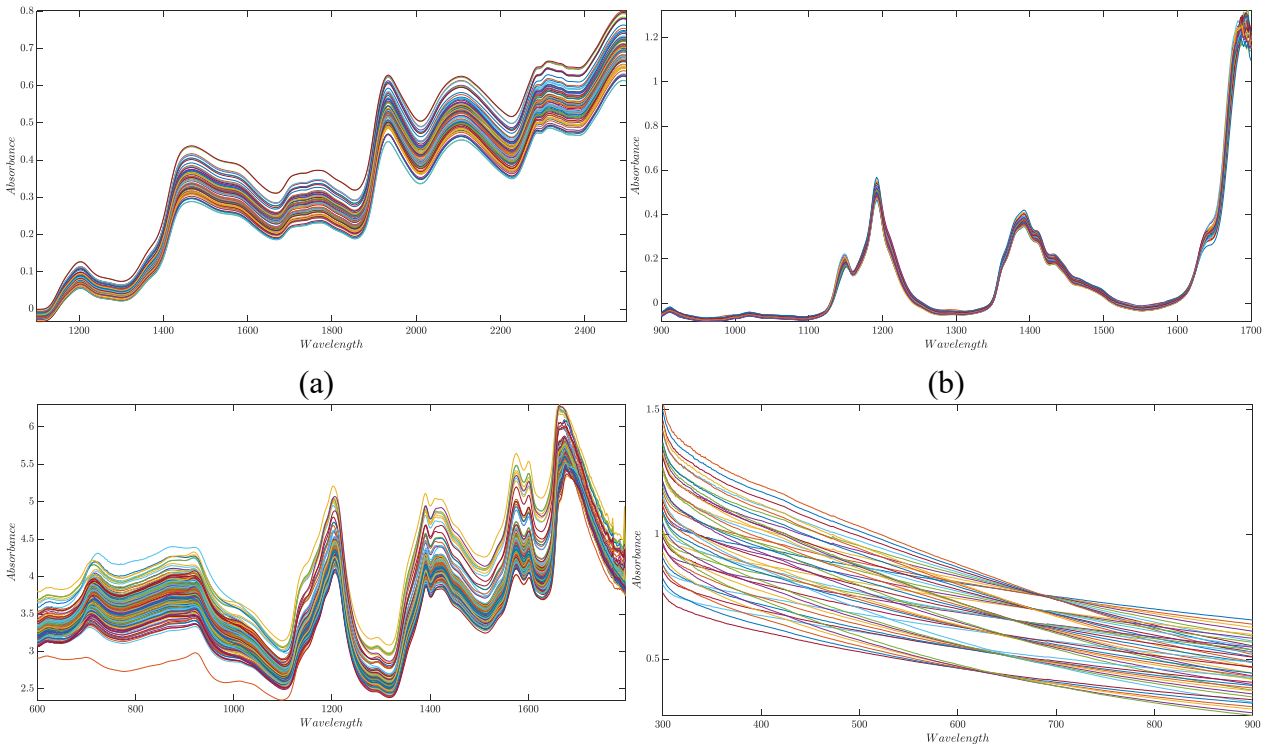


Figure 3. NIR spectra of (a) corn, (b) gasoline, (c) pharmaceutical tablet and (d) coculture datasets (different colors refer to different samples)

pharmaceutical dataset that has 655 samples, ~30% of all samples were left out as test samples while ~70% of all samples were used for calibration. In addition, to avoid model overfitting, based on the recommendation suggested by [47], about 45% of all calibration samples were left out for validation (i.e., data used for the optimization of the model parameters) while the rest of the calibration samples (about 55% of all calibration samples) were used for training (i.e., data actually used to build the models). Details of the data division for all datasets are given in Table 1. Literature [34] and our experiences suggest that such division of training and validation (i.e., ~55% vs. ~45%) results in models that are generally without overfitting issues. This is confirmed by the results of this work as discussed in details in Section 6. In addition, for small datasets such as most spectrum based datasets, the performance could be significantly affected by the data division (e.g., how many samples in training and testing respectively, and specific samples included in each group). To address this potential bias, we propose a Monte Carlo validation and testing (MCVT) procedure such that multiple (25 in this work) training and testing sets are randomly selected in each MC run and the average and standard deviation of the performances across different MC runs are used to robustly and fairly evaluate the soft sensor performance. Details are provided in the next section.

Table 1 Division of data into training, validation and test subsets

Dataset	Training (%)	Validation (%)	Test (%)	Total (%)
Corn	36 (45.0%)	28 (35.0%)	16 (20.0%)	80 (100%)
Gasoline	27 (45.0%)	21 (35.0%)	12 (20.0%)	60 (100%)
Pharma	263 (40.2%)	196 (29.9%)	196 (29.9%)	655 (100%)
Co-culture	21 (44.7%)	16 (34.0%)	10 (21.3%)	47 (100%)

6 Monte Carlo validation and testing (MCVT) procedure and MCVT-based performance indices

To systematically test the proposed method and compare its performance with full PLS, Lasso and SiPLS models, a Monte Carlo validation and testing (MCVT) procedure is proposed, which is based on Monte Carlo cross-validation (MCCV) [34], but adapted for the purpose of comparing performances across different methods. The MCVT procedure is outlined in Figure 4. In each outer (*i.e.*, prediction) MC loop, the MC sampling approach is applied on the full dataset to partition it randomly into a combined training-validation set and a test set based on specified proportions (e.g., proportions in Table 1). In each inner cross-validation MC loop, the MC sampling is applied on the training-validation set to generate a training set and a validation set, again, based on specified proportions such as those given in Table 1. A soft sensor model is built based on the training set and the validation set is projected onto the model with different model parameters to obtain a series of performance indices (*i.e.,* normalized root mean squared errors ($NRMSE_V$) as defined in Eqn. 1 but they can be other indices) as a function of model parameters (MP's). Table 2 lists MP's to be optimized for each soft sensor method. It is worth noting that the spectrum partition/segmentation is an important parameter to be tuned or optimized for interval based methods SiPLS and SPA. Table 2 also indicates that the calibration processes of SiPLS and SPA are more computationally intensive than those of full PLS and Lasso based soft sensors. $NRMSE_V$ has a dimension of $p_1 \times p_2 \times \cdots \times p_n$ where p_i is the number of discrete values of model parameter

i to be evaluated. The inner MC loop is repeated M_V times (which is 100 in this work) to generate M_V $NRMSE_V$'s, which complete the inner MC loop. The $NRMSE_V$'s are averaged over the M_V MC runs to generate $\overline{NRMSE_V}$. The MP's (e.g., number of PC's, etc.) that result in the lowest $\overline{NRMSE_V}$ is used to build a prediction model using the combined calibration set. The test set is then projected onto the prediction model to generate the performance index $NRMSE_P$. The outer MC loop is repeated M_P times (which is 25 in this work), resulting in $M_P \times M_V$ inner (calibration) MC loops, to generate M_P $NRMSE_P$'s. The mean of $NRMSE_P$'s (*i.e.*, $\overline{NRMSE_P}$ as defined in Eqn. 3) then can be used to evaluate the accuracy of the method while the standard deviation of $NRMSE_P$'s (*i.e.*, σ_{NRMSE_P} as defined in Eqn. 4) can be used to assess the precision, or robustness/consistency of the method. Other performance indices can also be included, such as average normalized mean prediction error ($\overline{NMPE}$ as defined in Eqn. 6) for quantifying prediction bias.

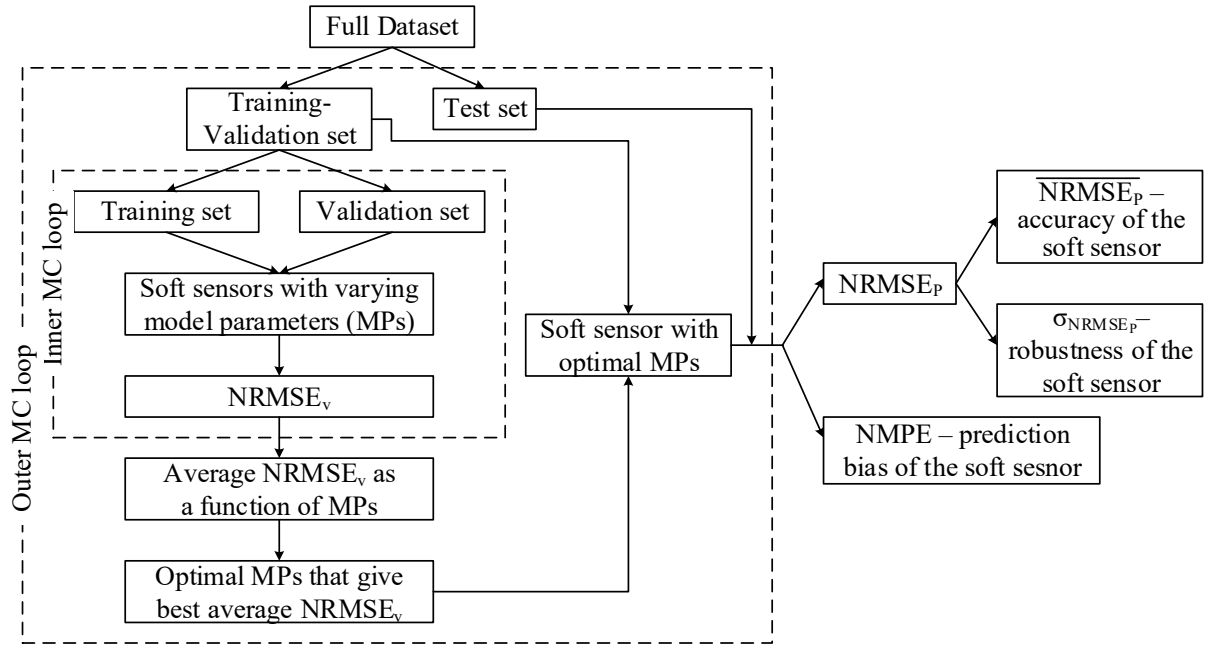


Figure 4. Flow diagram of the proposed Monte Carlo validation and testing procedure for comparing different soft sensor methods

Table 2. Parameters to be optimized for all methods

Method	**Parameters to be calibrated**	**No. of parameters to be calibrated**
Full PLS	No. of PC's	1
Lasso	λ	1
SiPLS	No. of segments the spectrum is divided into; Segments used in the model; No. of PC's	3
SPA (this work)	No. of segments the spectrum is divided into; Statistics/features used in the model; No. of PC's	3

As mentioned previously, the following three MCVT-based performance indices are proposed to evaluate the performance of each soft sensor in this work.

Normalized root mean squared error ($NRMSE$) as percentage of the measurement range:

$$NRMSE = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y-\hat{y})_i^2}}{(y_{max}-y_{min})} \times 100\% \quad (2)$$

Average $NRMSE$ ($\overline{NRMSE}$):

$$\overline{NRMSE} = \frac{\sum_{i=1}^{M} NRMSE_i}{M} \quad (3)$$

Standard deviation of NRMSE (σ_{NRMSE})

$$\sigma_{NRMSE} = \sqrt{\frac{\sum_{i=1}^{M}(NRMSE_i - \overline{NRMSE})}{M-1}} \quad (4)$$

Normalized mean prediction error (NMPE) as percentage of the measurement mean:

$$NMPE = \frac{\sum_{i=1}^{n}(y-\hat{y})_i}{\sum_{i=1}^{n} y_i} \times 100\% \quad (5)$$

Average $NMPE$ ($\overline{NMPE}$):

$$\overline{NMPE} = \frac{\sum_{i=1}^{M} NMPE_i}{M} \quad (6)$$

where n is the total number of validation (n_V) or prediction (n_P) samples in each MC run, and M is the total number of MC runs during validation (M_V) or prediction (M_P).

7 Results and Discussion:

All analyses follow the MCVT procedure discussed in the previous section (with $M_V = 100$ and $M_P = 25$) and MCVT-based performance indices (*i.e.*, $\overline{NRMSE}$, $\sigma_{NRMSE}$and $\overline{NMPE}$) were compared among different methods. For the proposed SPA soft sensor, after optimization during calibration/validation, the statistics and features selected for different datasets are listed in Table 3. As can be seen from Table 3, not all statistics and features are selected for all datasets. This is because the features are selected to minimize $\overline{NRMSE}_V$, which is similar to variable selection, where there is a trade-off between information added by the feature and noise and/or bias added by the feature.

Table 3. Statistics and features selected for different datasets

Dataset	Statistics and features selected
Corn dataset	μ, γ, κ, ASD, SLL, SSL
Gasoline dataset	μ, SLL
Pharmaceutical dataset	μ, γ, κ, AFD, ASD, SLL, SSL
Co-culture dataset	μ, σ

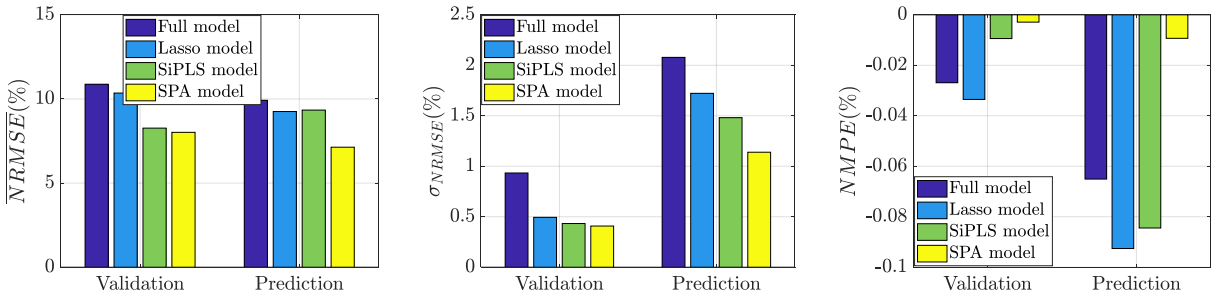
The average model sizes in terms of number of variables/features in the final soft sensor model over the 25 outer/prediction MC runs are listed in Table 4. It can be seen that all models with variable selection are substantially smaller than the full model. SPA has the smallest model size in three cases and the moderate model size in the rest two cases. In addition, it was found that the interval based methods, i.e., SiPLS and SPA, are quite robust to spectrum segmentation (i.e., the number of segments the spectrum is divided into). Due to limited space and the scope of this work, the results are not presented. In the following, we discuss the findings from the comparison of different methods on different datasets.

Table 4 Average number of variables/features of different soft sensors

Dataset	**PLS**	**SiPLS**	**Lasso**	**SPA**
Corn	700	152	141	84
Gasoline	401	84	14	30
Pharma	600	136	21	98
Co-cult (*E. coli*)	601	129	102	34
Co-cult (S. cerevisiae)	601	138	109	28

7.1 Comparison among full PLS, SiPLS, Lasso and SPA based soft sensors

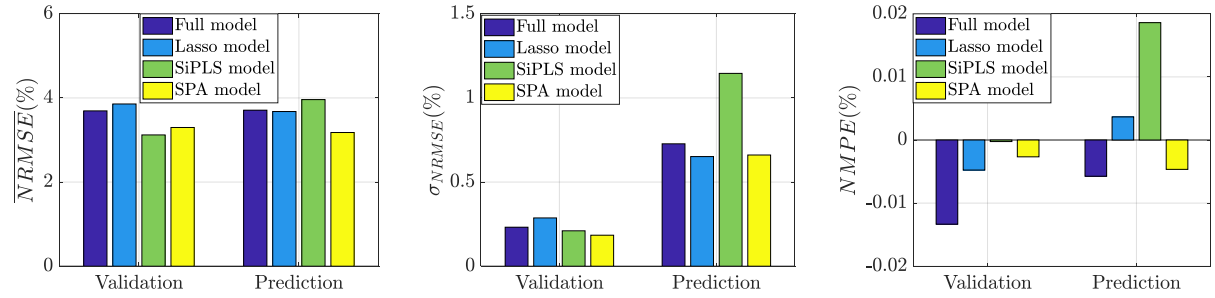
Corn data: Figure 5 shows comparison of all three indices, for validation and predication, obtained from the four approaches: full PLS, SiPLS, Lasso and SPA based soft sensors. First, if we compare performance indices (*i.e.*, $\overline{NRMSE}$, σ_{NRMSE}, and $\overline{NMPE}$) of validation vs. prediction, there is a general trend of performance deterioration from validation to prediction, which is expected as the model parameters were optimized based on the validation data. However, the performance deteriorations are not drastic, indicating that there is no obvious overfitting of models. The second observation is that in general variable selection improves soft sensor performance. Although $\overline{NMPE}_P$'s of Lasso and SiPLS are slightly higher than the full model in this particular case study, the error is insignificant – less than 0.1% of the measurement mean. The third observation is that for SiPLS, the model prediction performances are noticeably worse than that of the validation, especially σ_{RMSE}, and $\overline{NMPE}$. The likely reason is that the wavelengths selected by Lasso or SiPLS to optimize the prediction performance based on the calibration (*i.e.,* training and validation) data may be "tilted" to overfit or capture the noise or unknown disturbances contained in the calibration data. As a result, when the model is extrapolated or applied to new samples, the performance may deteriorate noticeably. Finally, SPA outperforms Lasso and SiPLS in all performance metrics. We believe the likely reason is that SPA does not discard any wavelength. Instead, SPA extracts features over the whole spectrum, making it more robust against performance degradation from validation to prediction. Overall, SPA-based soft sensor provides the best performance.



(a) (b) (c)

Figure 5. Comparison of soft sensors using corn data (moisture): (a) $\overline{NRMSE}$; (b) σ_{NRMSE}; (c) $\overline{NMPE}$

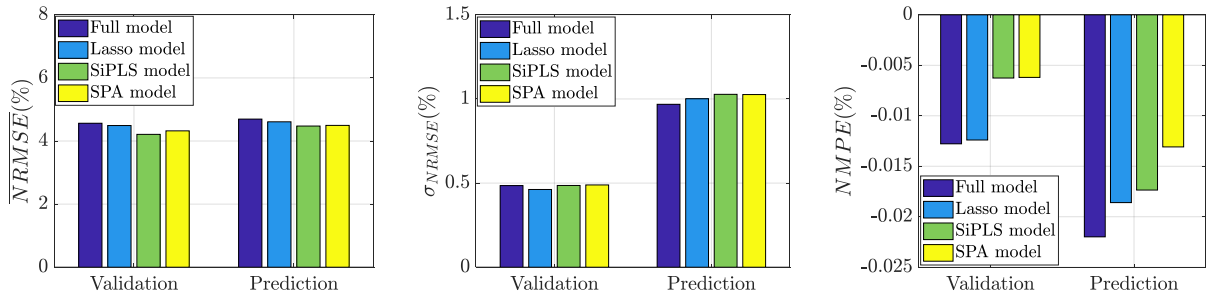
Gasoline data: As shown in Figure 6, for the gasoline data, similar trend of performance deterioration from validation to prediction is observed as expected. But again, the insignificant difference in $\overline{NRMSE}$ suggests no obvious overfitting of the models. In addition, the performance of SiPLS deteriorate noticeably from validation to prediction, especially σ_{NRMSE} and $\overline{NMPE}$, which increased by more than three folds and eight folds, respectively. In contrast, the performances of Lasso and SPA are more consistent with reasonable increase in σ_{RMSE} and $\overline{NMPE}$.



(a) (b) (c)

Figure 6. Comparison of soft sensors using gasoline data: (a) $\overline{NRMSE}$; (b) σ_{NRMSE}; (c) $\overline{NMPE}$

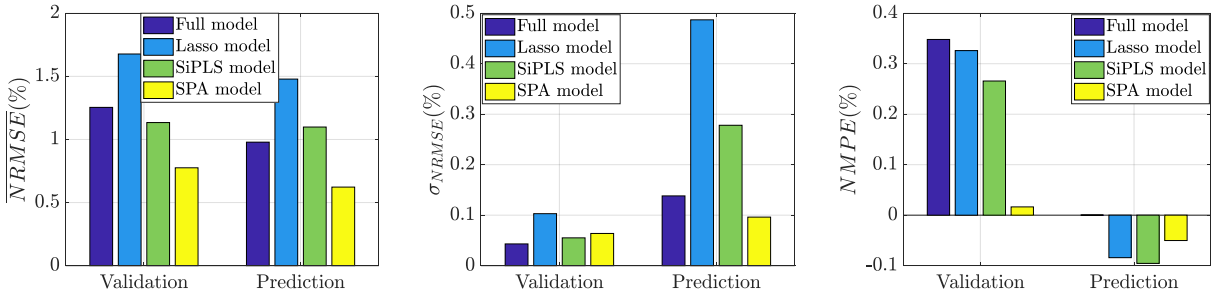
Pharmaceutical tablet data: As shown in Figure 7, for the pharmaceutical data, the performances of Lasso, SiPLS and SPA are similar, which are slightly better than that of the full model in terms of $\overline{NRMSE}$ and $\overline{NMPE}$ but with slightly higher σ_{NRMSE}. SPA gives the smallest $\overline{NMPE}$ among all four models.



(a) (b) (c)

Figure 7. Comparison of soft sensors using pharmaceutical data: (a) $\overline{NRMSE}$; (b) σ_{NRMSE}; (c) $\overline{NMPE}$

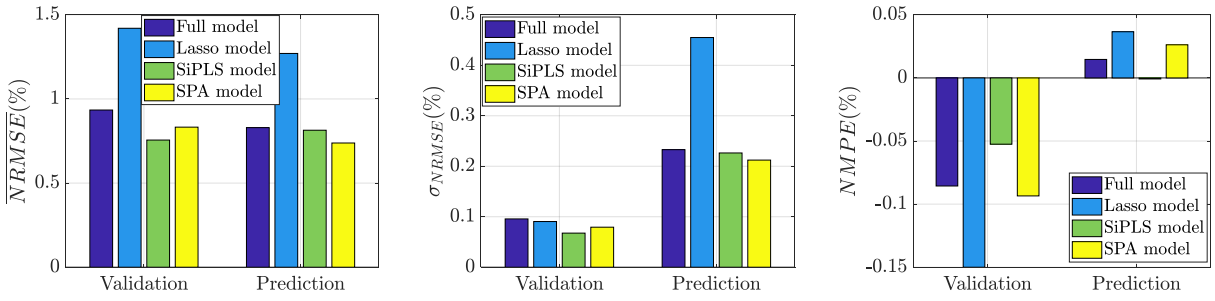
Co-culture data: For the coculture data, two sets of models are built to predict the concentrations of *E. coli* and *S. cerevisiae* separately. Separate models were used because the concentrations of the two species are properties that are not supposed to be correlated to each other. As shown in Figure 8, for *E. coli*, Lasso and SiPLS actually perform worse than the full model in prediction, but SPA performs noticeably better than the full model. Although the $\overline{NMPE}$ of SPA is higher than that of the full model, its value of less than 0.05% of the measurement mean indicates already very accurate predictions.



(a) (b) (c)

Figure 8 Comparison of soft sensors using coculture data (*E. coli*): (a) $\overline{NRMSE}$; (b) σ_{NRMSE}; (c) $\overline{NMPE}$

For *S. cerevisiae*, Lasso again performs worse than the full model, while SiPLS and SPA performing slightly better than the full model as shown in Figure 9. For $\overline{NMPE}$, again, the values from all models are less than 0.05% of the measurement mean indicates high accuracy of predictions from all models. When models for both species are considered, SPA performs better than Lasso and SiPLS. This further supports the SPA methodology of not removing any wavelength but rather extracting features from the full spectrum.

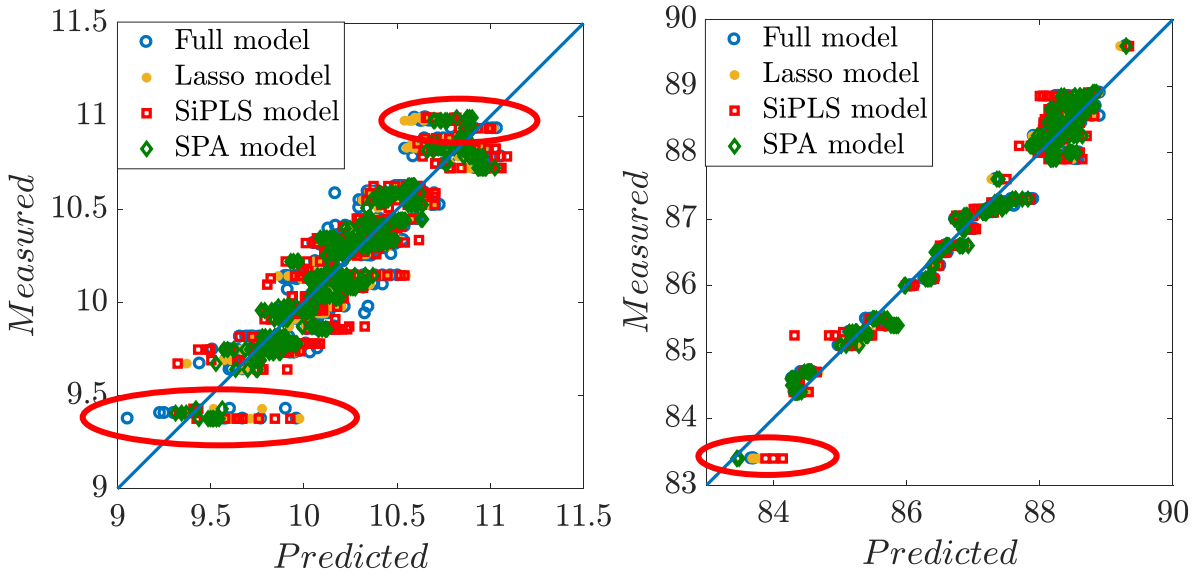


(a) (b) (c)

Figure 9. Comparison of soft sensors using coculture data (*S. cerevisiae*): (a) $\overline{NRMSE}$; (b) σ_{NRMSE}; (c) $\overline{NMPE}$

7.2 Discussions

In this section, we discuss potential reasons for the improved performance from SPA feature-based soft sensor. In addition, we examine the potential of KPLS based soft sensor applied to either original spectra or SPA features. For the corn data, Figure 10 (a) shows the predicted vs. measured moisture content, which confirms that SPA performs better than full PLS, Lasso and SiPLS across the whole property region. More importantly, we want to note that SPA performs especially better at extreme or boundary regions, as highlighted by red ellipses in Figure 10, where the number of samples are usually fewer than other regions. For example, when the moisture content is below 9.5, the predictions from the full PLS, Lasso and SiPLS models are widespread (*i.e.*, large σ_{NRMSE}) and with significant bias (*i.e.*, high $\overline{NMPE}$). Similar observations are found for the gasoline (Figure 10 (b)) and other datasets (not shown due to space limit). We postulate that the wavelengths selected by Lasso or SiPLS to optimize the performance based on the calibration data may be "tilted" to overfit or capture the noise or unknown disturbances contained in the calibration data, which are dominated by samples from the dense regions. As a result, the prediction performance of SiPLS may deteriorate significantly when the model is extrapolated or applied to samples from the sparse regions. As for the full PLS model, we believe the reason is because its performance is optimized by $NRMSE$, which means the spare regions with fewer samples would weigh less than the dense regions with more samples in determining the model parameters. As for SPA, although it is also optimized by $NRMSE$, it seems that the statistics and features extracted based on different regions of the full spectrum can alleviate this situation and the resulted model can extrapolate much better to the sparse regions with fewer samples than the models from PLS, Lasso and SiPLS.



(a) (b)

Figure 10. Comparison of predicted vs. measured properties from different soft sensors using (a) corn data and (b) gasoline data. Red ellipses highlight the regions where SPA performs significantly better than the full PLS, Lasso and SiPLS models.

In the previous section, we only compared linear soft sensors, although, strictly speaking, the mapping from the original spectrum to the SPA features is not linear. Here we explore the potential of KPLS as a nonlinear soft sensor for spectroscopic data analysis applications. For KPLS with a Gaussian kernel, the number of variables in the feature/kernel space equals the number of training samples, which is usually significantly larger than the number of variables under normal circumstances. However, this is not true for many spectroscopic data analysis applications as discussed previously. In those cases, KPLS actually shrinks the variable dimension in the kernel space. It is worth noting that this work is not intended to invalidate the merits of KPLS, but rather a case study to see if nonlinearity in the spectroscopic data can be captured for improving soft sensor performance given the severe constraint of the limited number of samples. In this work two scenarios are studied: the first scenario is to apply KPLS on the original full spectra; the second scenario is to apply KPLS on SPA features. The same MCVT procedure is followed to tune KPLS parameters, including the Gaussian kernel parameter σ and number of PC's. For the first scenario where KPLS is applied on the full spectra of each dataset, the perform of KPLS is poor for all four datasets. Table 5 compares the $\overline{NRMSE}$ and $\overline{NMPE}$ of 25 MC prediction runs for the pharmaceutical dataset, which is the largest dataset studied in this work with 263 training samples. Table 5 shows that KPLS performs even worse than full PLS. This is not surprising due to the severe constraint of limited number of samples. Therefore, the poor performance of KPLS does not indicate that there is no nonlinearity exists in the data, instead, it can only be said that KPLS cannot overcome the deficiency of variable shrinkage (instead of variable expansion in a regular KPLS application) due to the smaller number of samples (263) than that of variables (600). In the second scenario, we apply KPLS on top of SPA and term it SPA-KPLS. In other words, KPLS is replacing PLS and applied to SPA features. Again, the same MCVT procedure is followed to tune

the KPLS parameters. The performances of SPA-KPLS on the four datasets are compared to PLS-based SPA, and the results are listed in Table 6. Table 6 shows that KPLS does not improve the performance of SPA-based soft sensor in three out of four datasets (i.e., corn, gasoline and pharmaceutical datasets), which we believe can be attributed to the small number of training samples. However, KPLS does help in predicting *E. coli* and *S. cerevisiae* concentrations in the coculture dataset with only 21 samples for each strain. Although we do not have a definitive answer to explain this, we believe this is due to the significant similarity between the absorbance of the two strains, which makes the nonlinear interactions between the absorbance of the two strains an important factor in predicting their individual concentrations. In other words, the nonlinearity captured by KPLS outweighs the deficiency of dimension shrinkage due to limited samples.

Table 5 Prediction performance of KPLS on the pharmaceutical dataset compared to PLS and SPA

	PLS	**SPA**	**KPLS**
$\overline{\boldsymbol{NRMSE}}$	4.69%	4.49%	5.03%
$\overline{\boldsymbol{NMPE}}$	-0.022%	-0.013%	-0.032%

Table 6 Prediction performance of SPA-KPLS compared to SPA

		Corn	**Gasoline**	**Pharma.**	**E. coli**	**S. cerevisiae**
$\overline{\boldsymbol{NRMSE}}$	SPA	7.13%	3.18%	4.50%	0.62%	0.74%
	SPA-KPLS	8.04%	3.40%	4.73%	0.48%	0.44%
$\overline{\boldsymbol{NMPE}}$	SPA	-0.0093%	-0.0047%	-0.013%	-0.050%	0.026%
	SPA-KPLS	-0.055%	-0.011%	-0.0034%	-0.015%	0.033%

8 Conclusions and Future work

In conclusion, although variable selection in general could significantly improve soft sensor performance and reduce model complexity, there are potential pitfalls. As demonstrated by multiple case studies in this work, variable selection methods can be sensitive to the choice of training data and their performance could deteriorate noticeably when applied to test samples. We believe the possible reason is that the wavelengths selected (or wavelengths removed for that matter) to optimize the performance based on the calibration (*i.e.,* training and validation) data may be "tilted" to overfit or capture the noise or unknown disturbances contained in the calibration data. As a result, the model prediction performance may deteriorate significantly when the model is extrapolated or applied to new samples. To address this limitation, we propose a feature-based soft sensor approach by adapting the idea of SPA-based process monitoring framework we developed previously. Instead of selecting certain wavelengths or wavelength segments, the SPA feature-based soft sensor considers the whole spectrum, which is divided into segments, and extracts different features over each spectrum segment to build the soft sensor. In this way, there is no removal or exclusion of any wavelength or spectrum segment. As demonstrated in multiple case studies in this work, the proposed SPA feature-based soft sensor in general outperforms the original absorbance based soft sensor (*i.e.*, the full PLS soft sensor) as well as the variable selection

method Lasso or SiPLS based soft sensor in terms of $\overline{NRMSE}$, σ_{NRMSE}, and $\overline{NMPE}$. The SPA feature-based soft sensor is more robust than Lasso and SiPLS based soft sensor as evidenced by the smaller performance dip from validation to prediction. In addition, the SPA feature-based soft sensor can extrapolate much better to the sparse regions with fewer samples than the soft sensors based on full PLS, Lasso or SiPLS. We believe the main reasons for the good performance and robustness of the SPA based soft sensor are due to the following two factors: (1) features of spectrum segments correlate better to the property of interest (in a linear fashion through PLS) than the original spectrum or selected wavelengths; (2) inclusion of all information from the whole spectrum without removal or exclusion of any wavelength or wavelength segment enhances the robustness of the soft sensor. Finally, for small datasets such as most spectrum based datasets, the soft sensor performance could be significantly affected by the data division (e.g., how many samples in training and testing respectively, and specific samples included in each group). To address this potential bias, we propose a Monte Carlo validation and testing (MCVT) procedure such that multiple (25 in this work) training and testing sets are randomly selected in each MC run and the average and standard deviation of the performances across different MC runs are used to robustly and fairly evaluate the soft sensor performance across different datasets, which are generally applicable for model comparison in other applications. Although linear soft sensor methods are much preferred in most applications for their simplicity and interpretability, we tested the potential of nonlinear KPLS applied to both original spectra and SPA features. The results indicate that when the number of samples are severely limited, applying KPLS to full spectra is not a good solution compared to PLS. However, when KPLS is applied to SPA features, although still suffering the deficiency of small sample size, the results are much improved. The results of the coculture dataset justify the advantage of KPLS over PLS when there are potentially strong nonlinear interactions.

For future work, how to select features more systematically with less computation is worth further investigation. One potential route would be to integrate or adapt some variable selection approaches for feature selection. In addition, it is desirable if fundamental relations or theories could be developed for rational selection of features. Finally, integrating nonlinear regression methods, such as ANN and kernel-based approaches, into feature-based soft sensor is worth further investigation, especially for cases where large number of samples are available and/or potentially strong nonlinear interactions exist.

9 Acknowledgements

We acknowledge the constructive comments from the anonymous reviewers. Financial supports from National Science Foundation (NSF-CBET # 1805950) is greatly appreciated.

References:

[1] C.C. Felício, L.P. Brás, J.A. Lopes, L. Cabrita, J.C. Menezes, Comparison of PLS algorithms in gasoline and gas oil parameter monitoring with MIR and NIR, Chemom. Intell. Lab. Syst. 78 (2005) 74–80. doi:10.1016/j.chemolab.2004.12.009.

[2] A.M. Mouazen, J. De Baerdemaeker, H. Ramon, Towards development of on-line soil moisture content sensor using a fibre-type NIR spectrophotometer, Soil Tillage Res. 80

(2005) 171–183. doi:10.1016/j.still.2004.03.022.

[3] Q. Chen, J. Zhao, M. Liu, J. Cai, J. Liu, Determination of total polyphenols content in green tea using FT-NIR spectroscopy and different PLS algorithms, J. Pharm. Biomed. Anal. 46 (2008) 568–573. doi:10.1016/j.jpba.2007.10.031.

[4] K.A. Stone, D. Shah, M.H. Kim, N.R.M. Roberts, Q.P. He, J. Wang, A Novel Soft Sensor Approach for Estimating Individual Biomass in Mixed Cultures, Biotechnol. Prog. (2017).

[5] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemom. Intell. Lab. Syst. 2 (1987) 37–52. doi:10.1016/0169-7439(87)80084-9.

[6] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta. 185 (1986) 1–17. doi:10.1016/0003-2670(86)80028-9.

[7] T.-H. Pan, B.-Q. Sheng, D.S.-H. Wong, S.-S. Jang, A virtual metrology model based on recursive canonical variate analysis with applications to sputtering process, J. Process Control. 21 (2011) 830–839.

[8] D. Zhou, G. Li, S.J. Qin, Total projection to latent structures for process monitoring, AIChE J. 56 (2010) 168–178.

[9] F.A.A. Souza, R. Araújo, J. Mendes, Review of soft sensor methods for regression applications, Chemom. Intell. Lab. Syst. 152 (2016) 69–79. doi:10.1016/J.CHEMOLAB.2015.12.011.

[10] P. Kadlec, B. Gabrys, S. Strandt, Data-driven Soft Sensors in the process industry, Comput. Chem. Eng. 33 (2009) 795–814. doi:10.1016/j.compchemeng.2008.12.012.

[11] K. Bennett, M. Embrechts, An Optimization Perspective on Kernel Partial Least Squares Regression Neural networks View project Taguchi Methodology View project An Optimization Perspective on Kernel Partial Least Squares Regression, IOS Press Amsterdam, 2003. www.drugmining.com. (accessed February 5, 2019).

[12] R. Rosipal, L.J. Trejo, Kernel Partial Least Squares Regression in Reproducing Kernel Hilbert Space, J. Mach. Learn. Res. 2 (2001) 97–123. http://www.jmlr.org/papers/v2/rosipal01a.html (accessed February 5, 2019).

[13] V.N. Vapnik, The nature of statistical learning theory, Springer, 2000.

[14] M. Momma, K.P. Bennett, Sparse Kernel Partial Least Squares Regression, in: Springer, Berlin, Heidelberg, 2003: pp. 216–230. doi:10.1007/978-3-540-45167-9_17.

[15] Z. Wang, Q.P. He, J. Wang, Comparison of variable selection methods for PLS-based soft sensor modeling, J. Process Control. 26 (2015) 56–72. doi:10.1016/j.jprocont.2015.01.003.

[16] R.M. Balabin, S. V Smirnov, Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data, Anal. Chim. Acta. 692 (2011) 63–72.

[17] C.M. Andersen, R. Bro, Variable selection in regression---a tutorial, J. Chemom. 24 (2010) 728–737.

[18] X. Zou, J. Zaho, M.J.W. Povey, M. Holmes, H. Mao, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta. 667 (2010) 14–32.

[19] F. Lindgren, P. Geladi, S. Rännar, S. Wold, Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms, J. Chemom. 8 (1994) 349–363.

[20] I.-G. Chong, C.-H. Jun, Performance of some variable selection methods when multicollinearity is present, Chemom. Intell. Lab. Syst. 78 (2005) 103–112.

[21] E. Zamprogna, M. Barolo, D.E. Seborg, Optimal selection of soft sensor inputs for batch distillation columns using principal component analysis, J. Process Control. 15 (2005) 39–52.

[22] J. Liu, Developing a soft sensor based on sparse partial least squares with variable selection, J. Process Control. 24 (2014) 1046–1056.

[23] L. Cappellin, E. Aprea, P. Granitto, R. Wehrens, C. Soukoulis, R. Viola, T.D. Märk, F. Gasperi, F. Biasioli, Linking GC-MS and PTR-TOF-MS fingerprints of food samples, Chemom. Intell. Lab. Syst. 118 (2012) 301–307.

[24] C.-C. Pan, J. Bai, G.-K. Yang, D.S.-H. Wong, S.-S. Jang, An inferential modeling method using enumerative PLS based nonnegative garrote regression, J. Process Control. 22 (2012) 1637–1646.

[25] L. Nørgaard, A. Saudland, J. Wagner, J.P. Nielsen, L. Munck, S.B. Engelsen, Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy, Appl. Spectrosc. (2000). doi:10.1366/0003702001949500.

[26] G. Baffi, E.B. Martin, A.J. Morris, Non-linear projection to latent structures revisited: the quadratic PLS algorithm, Comput. Chem. Eng. 23 (1999) 395–411. doi:10.1016/S0098-1354(98)00283-X.

[27] A. Berglund, S. Wold, INLR, implicit non-linear latent variable regression, J. Chemom. 11 (1997) 141–156. doi:10.1002/(SICI)1099-128X(199703)11:2<141::AID-CEM461>3.0.CO;2-2.

[28] S. Wold, N. Kettaneh-Wold, B. Skagerberg, Nonlinear PLS modeling, Chemom. Intell. Lab. Syst. 7 (1989) 53–65. doi:10.1016/0169-7439(89)80111-X.

[29] S. Wold, Nonlinear partial least squares modelling II. Spline inner relation, Chemom. Intell. Lab. Syst. 14 (1992) 71–84. doi:10.1016/0169-7439(92)80093-J.

[30] R. Rosipal, Nonlinear partial least squares an overview, in: Chemoinformatics Adv. Mach. Learn. Perspect. Complex Comput. Methods Collab. Tech., IGI Global, 2011: pp. 169–189.

[31] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.

[32] J.-M. Lee, C. Yoo, S.W. Choi, P.A. Vanrolleghem, I.-B. Lee, Nonlinear process monitoring using kernel principal component analysis, Chem. Eng. Sci. 59 (2004) 223–

234.

[33] R. Tibshirani, Regression shrinkage and selection via the lasso, J. R. Stat. Soc. Ser. B. (1996) 267–288.

[34] M.H. Kim, Q.P. He, J. Wang, Quantifying the effects of oxygen utilization rate on ethanol production by S. stipitis under controlled chemostat, in: AIChE Annu. Conf., Salt Lake City, UT, 2015.

[35] L. Nørgaard, iToolbox Manual, 2005. www.models.kvl.dk.

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

[37] Q.P. He, J. Wang, Statistics Pattern Analysis - A New Process Monitoring Framework and Its Application to Semiconductor Batch Processes, AIChE J. 57 (2011) 107–121.

[38] Q.P. He, J. Wang, Statistics pattern analysis as a Big Data analytics tool for smart manufacturing, J. Process Control. 12 (2016).

[39] L.J. Janik, S.T. Forrester, A. Rawson, The prediction of soil chemical and physical properties from mid-infrared spectroscopy and combined partial least-squares regression and neural networks (PLS-NN) analysis, Chemom. Intell. Lab. Syst. 97 (2009) 179–188.

[40] H. Susi, D.M. Byler, Protein structure by Fourier transform infrared spectroscopy: Second derivative spectra, Biochem. Biophys. Res. Commun. 115 (1983) 391–397. doi:10.1016/0006-291x(83)91016-1.

[41] Y. de Micalizzi, First and second order derivative spectrophotometric determination of benzyl alcohol and diclofenac in pharmaceutical forms, Talanta. 47 (1998) 525–530. doi:10.1016/s0039-9140(98)00080-0.

[42] N. Zhao, Z. Wu, Q. Zhang, X. Shi, Q. Ma, Y. Qiao, Optimization of Parameter Selection for Partial Least Squares Model Development, Sci. Rep. 5 (2015). doi:10.1038/srep11647.

[43] Corn dataset. http://www.eigenvector.com/data/Corn/index.html.

[44] Gasoline dataset. https://www.mathworks.com/help/stats/sample-data-sets.html.

[45] Pharmaceutical dataset. http://www.idrc-chambersburg.org/shootout_2002.htm.

[46] D.W. Hopkins, Shoot-out 2002: transfer of calibration for content of active in a pharmaceutical tablet, NIR News. 14 (2003) 10–13.

[47] Q.-S. Xu, Y.-Z. Liang, Monte Carlo cross validation, Chemom. Intell. Lab. Syst. 56 (2001) 1–11. doi:10.1016/s0169-7439(00)00122-2.