Available online at www.sciencedirect.com

ScienceDirect

IFAC PapersOnLine 53-2 (2020) 11338–11343

Improving Featured-based Soft Sensing through Feature Selection

Jangwon Lee*, Jin Wang*, Jesus Flores-Cerrillo**, Q. Peter He*+

*Department of Chemical Engineering, Auburn University, AL 36849 USA
** Linde Digital, Linde plc, Tonawanda, NY 14150 USA
+ Corresponding Author (Tel: 334-844-7602; email: qhe@auburn.edu)

Abstract: Driven by the expanding applications of spectroscopic technologies, many advancements have been reported for soft sensor modeling, which infers a sample's properties from its spectroscopic readings. Because the number of wavelengths contained in a sample spectrum is usually much larger than the number of samples, "curse-of-dimensionality" is a common challenge that would affect the predictive power of the soft sensor. This challenge could be alleviated through variable selection. However, there is no guarantee that the truly relevant variables would be selected, and the selected variables are often (very) sensitive to the choice of training and validation data. To help address this challenge, we have developed a feature-based soft sensing approach by adapting the statistics pattern analysis (SPA) framework. In the SPA feature-based soft sensing, the features extracted from different segments of the complete spectrum were utilized to build the model. In this way, the information contained in the whole spectrum is used to build the model, while the number of the variables is significantly reduced. In this work, by integrating a variable selection approach we developed recently with SPA, we not only further improve the soft sensor's prediction performance, but also identify the key underlying chemical information from spectroscopic data. The performance of the improved feature-based soft sensing approach, termed SPA-CEEVS, is demonstrated using two NIR datasets, and compared with several existing soft sensing approaches.

Keywords: Soft sensor, Variable selection, Consistency enhanced evolution for variable selection (CEEVS), Statistics pattern analysis (SPA), NIR

1. INTRODUCTION

Soft sensors, which correlate the spectroscopic reading of a sample to its properties, offer a non-invasive, fast and inexpensive way to estimate the sample properties of interest. Due to these advantages, spectroscopic-based soft sensors have been successfully applied to many different fields, including agriculture, pharmaceutical, oil and gas industries. Among many different modeling approaches, partial least squares (PLS) is the most commonly used multivariate statistical method, due to its simplicity, robustness and inherent capability in handling collinearity among regressors (Geladi & Kowalski, 1986).

For spectroscopic measurements, each sample spectrum contains hundreds or thousands of wavelengths (variables), and readings from adjacent wavelengths are usually highly correlated. However, most spectroscopic datasets contain rather limited number of samples, usually less than 100. It is well recognized that PLS works well when the number of samples is 20 time more than the number of variables. Clearly, this is not the case for the spectroscopic datasets. Variable selection could offer a potential solution to the problem, as readings from adjacent wavelengths are often (highly) correlated and not all spectrum segments are informative. As a result, variable selection has drawn significant research interest for soft sensor development, particularly for spectroscopic-based soft sensors (Balabin & Smirnov, 2011; Z. Wang, He, & Wang, 2015).

Variable selection has enjoyed many successful applications to improve soft sensor prediction, but it does have limitations. Specifically, the selected variables, hence the resulted soft sensor model, can be highly sensitive to the choice of training and validation data. Such sensitivity has been illustrated by the inconsistent variable selection results obtained from different Monte Carlo (MC) runs that randomly partitioning the data set into training and validation subsets, including the ones shown in this work. Due to the unknown disturbances and noises contained in the training and validation data, the soft sensor model may be "tilted" to overfit or to capture the unknown disturbance or noise contained in the training and validation set, and its performance could deteriorate significantly when applied to new samples.

To address this challenge, we have developed a feature-based soft sensor approach by adapting the basic idea of statistics pattern analysis (SPA) based process monitoring framework (Q. P. He & Wang, 2011; J. Wang & He, 2010). In the SPA-based soft sensor approach, instead of selecting certain wavelengths or wavelength segments, we make use of the whole sample spectrum. Specifically, the whole spectrum is divided into segments, and the selected features over each spectrum segment are used to build the soft sensor model (Shah, Wang, & He, 2019). In this way, the information contained in the whole spectrum is utilized but the number of variables used for model building is significantly reduced. As demonstrated in multiple case studies, SPA feature-based soft sensor in general outperforms the full PLS model that includes the whole spectrum, as well as PLS with variable selection, such as Lasso and SiPLS (Shah et al., 2019).

10.1016/j.ifacol.2020.12.542

However, it has been well-recognized that not all wavelengths (or wavelength segments) contribute equally to the sample property at interest. Because the sample property at interest is usually determined by certain chemical bonds or functional groups of the sample, only those absorption peak/valley corresponding to the chemical bonds or functional groups are the truly relevant inputs. Therefore, if the truly relevant spectrum segments could be selected for model building, variable selection would be highly desirable. To this end, we have developed a new variable selection method based on Darwin's evolution theory, i.e., "survival of the fittest". The new variable selection method is termed consistency enhanced evolution for variable selection (CEEVS), which focus on improving the consistency of variable selection results from different training datasets. We hypothesize that improved variable selection consistency would result in improved prediction performance. This is because the truly relevant input variables stay the same regardless of the choice of the training dataset. If a variable selection method can consistently select a subset of variables, it is likely that the selected ones are the truly relevant ones. Indeed, several case studies confirmed our hypothesis, and the wavelengths selected by CEEVS cluster around spectrum peaks and valleys which are associated with different chemical bonds and functional groups contained in the sample.

Compared to other variable selection methods that are based on Darwin's evolution theory, CEEVS shows better selection consistency, better model prediction performance. CEEVS usually has more variables being selected because CEEVS select the segments of the wavelengths clustered around peaks and valleys. In (Lee, Flores-Cerrillo, Wang, & He, 2020) we have verified that the wavelength segments selected by CEEVS indeed reveal underlying chemical information, as they correspond to different chemical bonds or functional groups. Although the wavelengths around peaks/valleys are highly correlated, all of them being consistently selected suggested that the shape (or area) of the peak, in additional to the height of the peak, are important to predict the sample properties. If this is the case, then the features extracted from the wavelength segments could provide the same information as all the wavelength together, which could provide same or even better prediction performance, while significantly reduce the number of the variables. Therefore, we apply CEEVS to select relevant features used in the SPA feature-based soft sensor, and examine its performance by comparing with existing methods.

The rest of the paper is organized as the follows. Sections 2 and 3 briefly introduce the SPA featured-based soft sensor framework and CEEVS, respectively; Section 4 presents the proposed SPA-CEEVS and Section 5 uses two case studies to demonstrate its performance, which is compared with SPA, CEEVS, as well as three representative variable selection methods that are based on the "survival of the fittest" principle; Section 6 draws conclusion.

2. SPA FEATURE-BASED SOFT SENSOR

Statistics pattern analysis (SPA) is a process monitoring framework that the authors developed previously (Q. P. P. He & Wang, 2018; Q. P. He & Wang, 2011; J. Wang & He, 2010), in which the statistics of process variables, instead of the process variables themselves, are monitored to determine the process operation status. Its effectiveness and performance in process monitoring have been demonstrated in multiple case studies (Q. P. P. He & Wang, 2018; Q. P. He & Wang, 2011; J. Wang & He, 2010). In the original SPA based process monitoring approach, the statistics are calculated along the time dimension and principal component analysis (PCA) is performed on the statistics for fault detection and diagnosis. In the 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) (*i.e.*, sample properties) through PLS. The schematic diagram of the SPA feature-based soft sensor approach is shown in Figure 1, where we first divide the whole sample spectrum into s non-overlapping segments; then f different features are extracted from each spectrum segment. 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. In this way, information from the whole spectrum will be utilized for model building, but with significantly reduced number of variables.

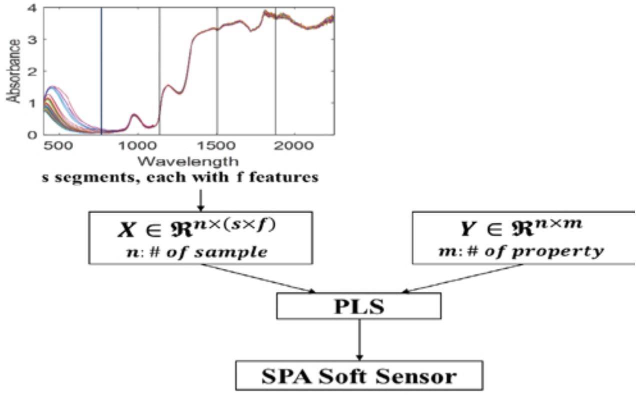


Fig. 1. Schematic of SPA feature-based soft sensor

More details about SPA feature-based soft sensor, as well as its performance when compared to other methods on multiple case studies, can be found in (Shah et al., 2019).

3. CONSISTENCY ENHANCED EVOLUTION FOR VARIABLE SELECTION (CEEVS)

It is clear that the truly relevant input variables would stay the same regardless of the choice of the training data set. Therefore, in CEEVS, we focus on improving the consistency of variable selection results from different training datasets. We hypothesize that better variable selection consistency would result in better soft sensor prediction performance, because if a variable is selected consistently across different training sets, it is more like a truly relevantly regressor.

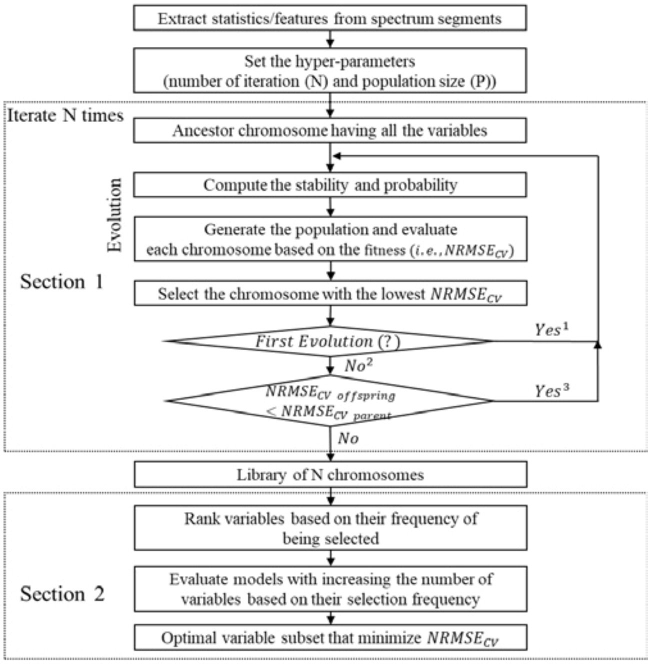
The CEEVS method is also based on the "survival of the fittest" principle, and follows the same terminologies as genetic algorithm (GA). A gene refers to an individual variable, and a chromosome ($c_{p\times1}$) refers to a set of selected variables. For example, the i-th element (c_i) of the chromosome, either "1" or "0", indicates whether the i-th wavelength is included in the chromosome or not, respectively. In CEEVS, we rely on random MC sampling of the sample space to assess the stability of each variable, which is determined based on how consistently the variable contributes to the soft sensor model derived from different training samples. This stability is then converted into "probability for selection", based on which the initial "chromosome" population will be generated.

As shown in Fig. 2, CEEVS consists of two main sections. In Section 1, starting with the complete variable set, the initial chromosome population is generated based on each variable's probability for selection. In this way, the evolution process will start with a better initial population, as more important variables will more likely be selected for the initial population. Then each chromosome is evaluated for its fitness value. The selected variables (*i.e.*, the variables that have "1" in the chromosome) are used to build a PLS model, and chromosome's fitness value is defined as the model's normalized root mean square error from cross-validation ($NRMSE_{CV}$). The optimal chromosome, *i.e.*, the one with the minimal $NRMSE_{CV}$ within the initial populations, is used as a parent to generate offspring for the evolution process.

The objective of the evolution process is to further eliminate the uninformative variables in the parent chromosome before it is stored into the library. During the evolution process, instead of cross-over and mutation, the variables selected by the parent chromosome are used as the new full variable set, and repeat the whole process to generate the next best chromosome which is denoted as an offspring. For each additional run of evolution, the offspring from the previous run is used as the parent chromosome to generate new offspring. In this way, all the offspring are guaranteed to contain fewer variables than the parent and have a better fitness value. This evolution process is repeated until the fitness of the offspring is worse than that of the parent, then the parent of the final evolution run, *i.e.*, the best chromosome generated from the whole evolution process, is stored into the library. This evolution process will be repeated N times, and each time starting with the complete set of variables. At the end of N iterations, the library will contain N optimally evolved chromosomes, i.e., subsets of selected variables that deliver the lowest $NRMSE_{CV}$ during each evolution process.

In Section 2, starting with the library that contains N best chromosomes, we first rank all the variables based on their frequency of presence in the library. Next, we build a series of PLS models with increasing number of variables based on their selection frequency. In other words, the first PLS model is built with the most frequently selected variables in the library and the second model adds the next frequently selected variable. This process is repeated until the number of variables included in the model reaches a pre-defined upper limit, which can be adjusted to reduce the risk of overfitting. In this work, we set the upper limit as 300. Finally, all models are evaluated for their fitness ($NRMSE_{CV}$), and the variable subset that produce the lowest $NRMSE_{CV}$ value is considered the final result of the selected variables.

More details about CEEVS can be found in (Lee et al., 2020), where CEEVS was tested with 5 different case studies. In addition, CEEVS was compared with 3 representative variable selectin methods that are also based on the "survival of the fittest" principle: genetic algorithm (GA) (Leardi, 2000; Leardi & Lupiáñez González, 1998), competitive adaptive reweighted sampling (CARS) (Li, Liang, Xu, & Cao, 2009) and stability and variable permutation (SVP) (Chen, Yang, Zhu, Li, & Gui, 2018). We confirmed that through enhancing variable selecting consistency, CEEVS delivers the best prediction performance. More importantly, we demonstrated that CEEVS is able to identify the underlying chemical information contained in the spectrum, *i.e.*, the key chemical bonds or functional groups that determine the sample property of interest.


Extract statistics/features from spectrum segments
Set the hyper-parameters
(number of iteration (N) and population size (P))
Iterate N times
Ancestor chromosome having all the variables
Evolution
Compute the stability and probability
Generate the population and evaluate
each chromosome based on the fitness (i.e., NRMSE_CV)
Section 1
Select the chromosome with the lowest NRMSE_CV
First Evolution (?)
Yes¹
No²
NRMSE_CV offspring < NRMSE_CV parent
Yes³
No
Library of N chromosomes
Rank variables based on their frequency of being selected
Section 2
Evaluate models with increasing the number of variables based on their selection frequency
Optimal variable subset that minimize NRMSE_CV

1. The selected chromosome to be parent chromosome.
2. The selected chromosome in previous evolution to be parent chromosome.
   The selected chromosome in current evolution to be offspring chromosome.
3. The offspring to be parent chromosome.

Fig. 2 Flow diagram of SPA-CEEVS algorithm

4. SPA FEATURE-BAED SOFT SENSING INTEGRATED WITH CEEVS (SPA-CEEVS)

In (Lee et al., 2020), we also found that CEEVS usually select the largest number of wavelengths, and the selected wavelengths consistently cluster around spectrum peak or valleys, which is how the underlying chemical information is identified. This makes sense, because the general features of molecular spectra are of continuous bands, and the shape of the peak or valley, in addition to peak height, could contain important information about the underlying molecular structure. As the shape of the peak cannot be captured by a single wavelength, this is why a segment of wavelengths around a peak or valley were consistently selected by CEEVS. However, the wavelengths within the peak/valley segment are highly correlated and do contain many redundant information. If such information could be captured by different features, we don't have to include the whole segment of the wavelengths, therefore reducing the number of regressors without scarifying prediction performance. In this work, we propose to integrate

SPA feature-based soft sensor with CEEVS for feature selection to simplify the soft sensor model.

In SPA-CEEVS, rooted in SPA feature-based soft sensing, we apply CEEVS to select relevant features, which are then used to build the soft sensor model. In this way, we could obtain a significantly simplified model while maintaining sensor performance, as we will use a few features to capture the key information contained in a spectrum segment; in addition, we could further enhance the prediction performance, as irrelevant features are removed through feature selection. Finally, the key chemical information could be identified through feature selection, similar to CEEVS.

5. CASE STUDIES

In this work, we use two published NIR datasets to illustrate the performance of the SPA-CEEVS method and compare its performance with that of SPA and CEEVS. In addition, the full PLS model that uses all the wavelengths as the regressors is provided as baseline, plus two representative variable selection methods, *i.e.*, GA, CARS, for comparison. In all methods, the soft sensor is constructed using PLS, either with all variables (full PLS model), or with selected variables (GA, CARS and CEEVS), or with full features based on full spectrum (SPA), or with selected features based on full spectrum (SPA-CEEVS).

Table 1 summarizes the two datasets, including the number of samples and variables, the partition of the dataset into training and testing subsets, as well as relevant references.

Table 1. Summary of the five NIR datasets

	# of calibration samples	# of test samples	# of variables	Property of interest	Reference
Beer	48 (80%)	12 (20%)	926	Extract concentration	(Nørgaard et al., 2000)
Pharma	459 (70%)	196 (30%)	650	Active pharmaceutical ingredients (API)	(Pharma. dataset)

To eliminate the potential bias caused by a specific partition of the whole dataset into calibration and testing subsets, we conduct 100 MC runs and use the results from all MC runs to evaluate the performance of each variable selection method. For each MC run, the calibration and testing subsets are randomly selected according to the percentage listed in Table 1. The normalized root mean square error in prediction ($NRMSE_P$) as defined below is used to evaluate the soft sensor prediction performance.

$$NRMSE_P = \frac{\sqrt{\frac{1}{N_T}\sum_{i=1}^{N_T}(y_i-\hat{y}_i)^2}}{(y_{max}-y_{min})} \times 100\% \tag{9}$$

where N_T is the number of test samples in each MC run. The normalization in $NRMSE_P$ facilitates the comparison of different methods across different datasets. The mean and the standard deviation of $NRMSE_P$ obtained from the 100 MC runs are used as the two metrics to evaluate the prediction performance of the soft sensor models. The mean ($\overline{NRMSE_P}$) can be used to evaluate the accuracy of each method while the standard deviation (σ_{NRMSE_P}) can be used to assess the robustness of the method. To directly measure the consistency of the variable selection among 100 MC runs, we use the following consistency index (I_c) (Lee et al., 2020).

$$I_c = \frac{\sum_{i=1}^{p} prob(x_i)}{m} \tag{10}$$

where m is the number of variables being selected at least once among all MC runs; $prob(x_i)$ is the probability of i^{th} variable being selected, which is defined as the ratio of selection frequency of i^{th} variable among all MC runs to number of MC runs. A higher I_c represents a better consistency, which indicates the informative variables are being more consistently selected regardless of calibration datasets.

To fairly compare different variable selection methods, each method is optimized based on 10-fold cross-validation. An exhaustive search is used to determine the optimal tuning parameters for each method.

5.1 Performance comparison

The results from all the methods are summarized in Tables 2 and 3. The best performance for each metric is represented in boldface. The improvement rate (%) indicates the improvement of $\overline{NRMSE_P}$, compared to the full PLS model; n_{PC} is the number of principal components in the model; n_{Var} is the number of selected variables in the final model.

Table 2. Performance comparison for the beer dataset

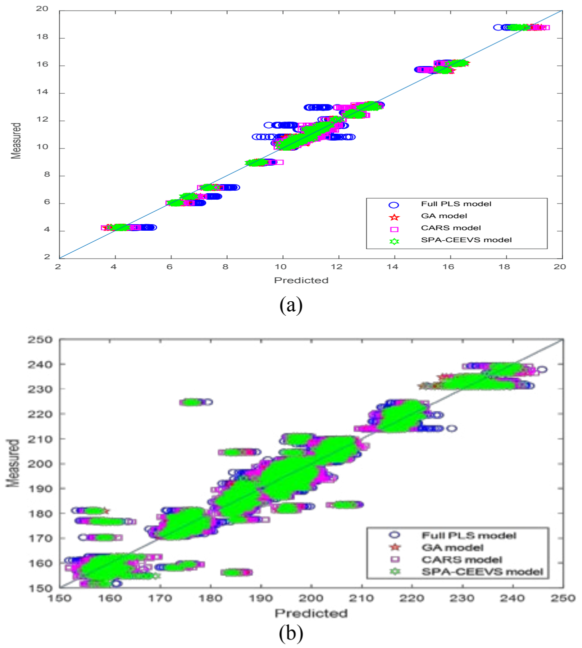
Method	$\overline{NRMSE_P}$	σ_{NRMSE_P}	Ic	Improv. (%)	n_{PC}	n_{Var}
Full PLS	6.57	6.46	-	-	9 ± 3	926
GA	2.37	1.85	0.142	63.91	8 ± 3	94 ± 58
CARS	3.24	2.76	0.192	50.64	9 ± 3	87 ± 38
CEEVS	2.36	1.45	0.182	64.11	8 ± 3	130 ± 86
SPA	3.22	2.40	-	50.98	8 ± 3	104
SPA-CEEVS	**1.77**	**1.21**	**0.249**	**73.07**	8 ± 3	**14 ± 7**

Table 3. Performance comparison for the pharmaceutical dataset

Method	$\overline{NRMSE_P}$	σ_{NRMSE_P}	Ic	Improv. (%)	n_{PC}	n_{Var}
Full PLS	5.05	**0.76**	-	-	14 ± 3	650
GA	4.46	0.90	0.138	11.69	11 ± 3	69 ± 44
CARS	4.72	0.84	0.064	6.50	15 ± 3	30 ± 15
CEEVS	4.45	0.89	0.231	11.86	13 ± 2	92 ± 56
SPA	4.53	0.88	-	10.28	10 ± 3	128
SPA-CEEVS	**4.43**	0.89	**0.338**	**12.15**	13 ± 3	**27 ± 9**

As shown in the tables, for both case studies, SPA-CEEVS offers the best prediction performance and the best selection consistency, as well as the simplest model with the smallest number of variables included. For the pharmaceutical dataset, although the full PLS model has the smallest standard deviation of NRMSE, the mean of the NRMSE is significantly larger than that of the other methods.

Fig. 4 provides the detailed comparison of the prediction performance from selected methods (Full PLS, GA, CARS and SPA-CEEVS) for 100 MC runs. As shown in Fig. 4, the predicted values by SPA-CEEVS clustered the closest to the diagonal line given different training data, demonstrating superior prediction accuracy and robustness than other methods.



(a)

(b)

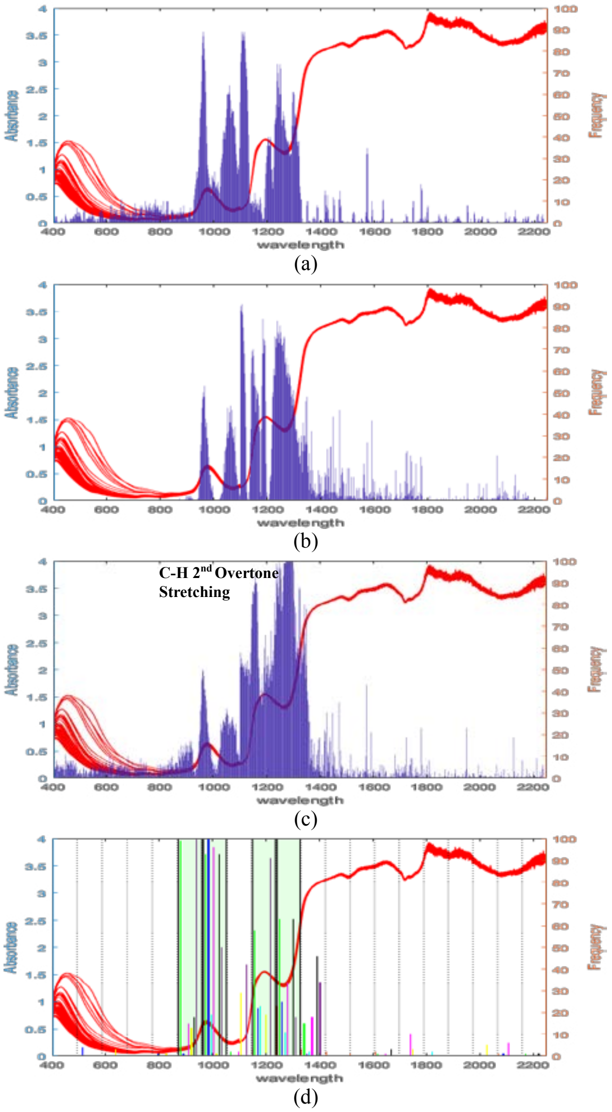
Fig. 4. Predicted vs. measured properties from all methods. (a) Beer dataset; (b) Pharmaceutical dataset.

5.2 Discussion

From Tables 2 and 3, it is interesting to notice that although the number of selected input variables varies significantly for different soft sensor models, the number of PCs selected by each soft sensor are very close to each other. The consistent number of PCs from different soft sensor models for each dataset suggests that the interdependence between the absorption spectrum and the sample properties is relatively simple and likely nonlinear, which is why large number of wavelengths were selected by different variable selection methods to achieve their corresponding optimal prediction performance. Because SPA use features that could directly capture nonlinear characteristics as input variables to build the model, the selected number of features is much smaller than that of absorbance-based soft sensors.

The superior prediction performance, both accuracy and robustness, by SPA-CEEVS can be contributed to two factors: first is that features, especially the nonlinear ones, could be more effective in capturing the underlying nonlinear relationship between sample spectrum and property of interest. Because PLS only captures linear relationship between input and output variables, such nonlinear relationship can only be linearly approximated by including larger number of absorbances at different wavelengths to balance out their nonlinear effects. Second, when a segment of wavelengths are used to compute different features, there is a built-in effect of noise filtering. For example, when the mean or standard deviation is computed, it is obtained as an average over the wavelength segment, therefore reduces the effect of potential noise contained in the absorbance spectrum. Finally, if only truly relevant variables are included, it is expected to deliver more accurate and robust prediction performance.

One major advantage of CEEVS is that it could reveal the underlying chemical bonds or functional groups by selecting relevant variables consistently. To examine whether this property is conserved for SPA-CEEVS, we plotted the variable selection frequency from different methods among 100 MC simulations (Fig. 5 and Fig. 6). For the beer dataset, the wavelength segment consistently selected by SPA-CEEVS agree with that selected by CEEVS, and it did not select any features corresponding to the initial noisy segment (400 – 800nm). For the pharmaceutical dataset, SPA-CEEVS covers wider wavelength segments than CEEVS, with more features selected for the segments selected by CEEVS. This suggests that SPA-CEEVS could also identify the key underlying chemical information in the sample spectrum.



(a)

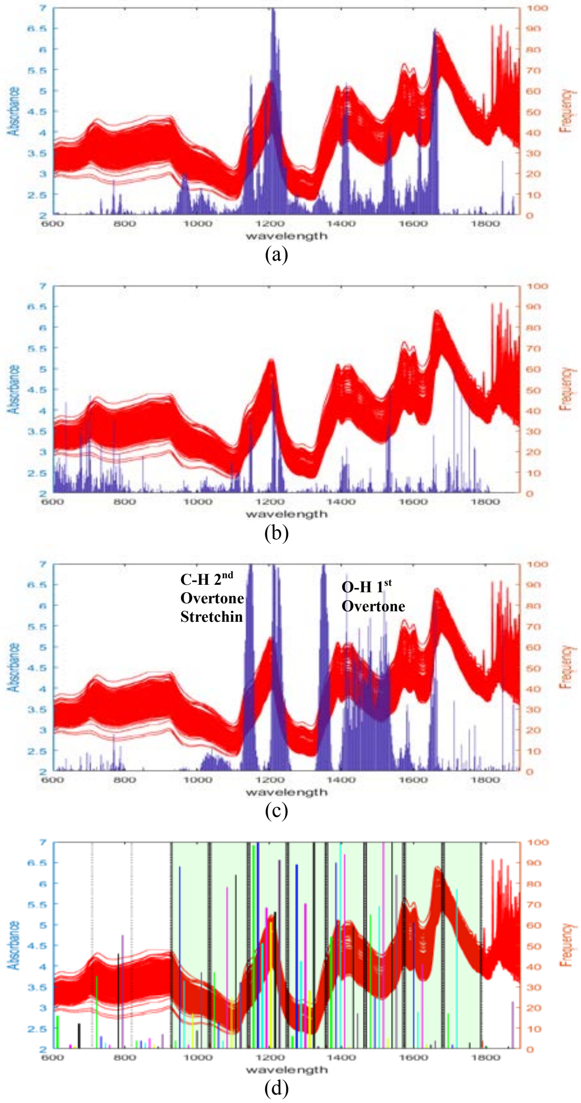
(b)

(c)

(d)

Fig. 5. Plot of spectra and selected variables over 100 MC runs for the Beer dataset. (a) GA; (b) CARS; (c) CEEVS; (d) SPA-CEEVS. In the SPA-CEEVS, the bars with different colors correspond to different statistics (brown: μ, green: σ, blue: γ, bright blue: κ, pink: AFD, yellow: ASD, black: SLL, purple: SSL). The dotted line denotes each segment.

When we compare the performances between SPA and SPA-CEEVS, we see that SPA-CEEVS can provide further improvement. This is because not all features of all wavelength segments contribute equally to the sample properties. With CEEVS to remove irrelevant features, SPA-CEEVS could further improve the prediction performance, while potentially identify the chemical bonds or functional groups that determine the sample property.



(a)

(b)

(c)

(d)

Fig. 6. Plot of spectra and selected variables over 100 MC runs for the Pharmaceutical tablet dataset. (a) GA; (b) CARS; (c) CEEVS; (d) SPA-CEEVS. In the SPA-CEEVS, the bars with different colors correspond to different statistics same as in Fig. 5. The dotted line denotes each segment.

6. CONCLUSION

Variable selection for soft sensor development has drawn significant research interest recently, driven by the application of spectroscopic soft sensors in different industries. However, one unsolved challenge is that the selected variables can be highly sensitive to the choice of training data, and may not be truly relevant variables. To address this challenge, we have previously developed a SPA feature-based soft sensing framework that use extracted features from sample spectrum to build the model, and a consistency enhanced evolution for variable selection (CEEVS) that have been shown to be able to identify underlying chemical information directly related to the sample property. In this work, we integrate CEEVS with SPA feature-based soft sensor, and demonstrate that the integrated approach, SPA-CEEVS, not only results in significantly simplified model and further improved prediction performance, but also could identify key underlying chemical information.

7. ACKNOWLEDGEMENT

Financial supports from National Science Foundation, NSF-CBET #1805950 (Lee, Flores-Cerrillo, Wang and He), and U.S. Department of Energy, DE-SC0019181 (Wang and He) are greatly appreciated.

REFERENCES

Balabin, R. M., & Smirnov, S. V. (2011). Variable selection in near-infrared spectroscopy: benchmarking of feature selection methods on biodiesel data. *Analytica Chimica Acta*, *692*(1), 63–72.

Chen, J., Yang, C., Zhu, H., Li, Y., & Gui, W. (2018). A novel variable selection method based on stability and variable permutation for multivariate calibration. *Chemometrics and Intelligent Laboratory Systems*, *182*, 188–201.

Geladi, P., & Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, *185*(C), 1–17.

He, Q. P. P., & Wang, J. (2018). Statistical process monitoring as a big data analytics tool for smart manufacturing. *Journal of Process Control*, *67*, 35–43.

He, Q. P., & Wang, J. (2011). Statistics pattern analysis: A new process monitoring framework and its application to semiconductor batch processes. *AIChE Journal*, *57*(1), 107–121. https://doi.org/10.1002/aic.12247

Leardi, R. (2000). Application of genetic algorithm-PLS for feature selection in spectral data sets. In *Journal of Chemometrics* (Vol. 14, pp. 643–655).

Leardi, R., & Lupiáñez González, A. (1998). Genetic algorithms applied to feature selection in PLS regression: How and when to use them. *Chemometrics and Intelligent Laboratory Systems*, *41*(2), 195–207.

Lee, J., Flores-Cerrillo, J., Wang, J., & He, Q. P. (2020). Consistency-Enhanced Evolution for Variable Selection Can Identify Key Chemical Information from Spectroscopic Data. *Industrial & Engineering Chemistry Research*, *59*(8), 3446–3457.

Li, H., Liang, Y., Xu, Q., & Cao, D. (2009). Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta*, *648*(1), 77–84.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., Engelsen, S. B., … Engelsen, S. B. (2000). Interval partial least-squares regression (iPLS): a comparative chemometric study with an example from near-infrared spectroscopy. *Applied Spectroscopy*, *54*(3), 413–419.

Pharma. dataset. http://www.idrc-chambersburg.org/shootout_2002.html

Shah, D., Wang, J., & He, Q. P. (2019). A feature-based soft sensor for spectroscopic data analysis. *Journal of Process Control*, *78*, 98–107.

Wang, J., & He, Q. P. (2010). Multivariate Statistical Process Monitoring Based on Statistics Pattern Analysis. *Industrial & Engineering Chemistry Research*, *49*(17), 7858–7869.

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