Consistency-Enhanced Evolution for Variable Selection Can Identify Key Chemical Information from Spectroscopic Data

Jangwon Lee[1], Jesus Flores-Cerrillo[2], Jin Wang[1*], Q. Peter He[1*]

[1]Department of Chemical Engineering, Auburn University, Auburn, AL 36849 USA

[2]Linde Digital, Linde plc, Tonawanda, NY, 14150

*JW: wang@auburn.edu; QPH: qhe@auburn.edu

Abstract

In the last few decades, spectroscopic techniques such as near-infrared (NIR) spectroscopy have gained wide applications in several industries, such as pharmaceutical, agricultural, oil and gas industries. As a result, various soft sensors have been developed to predict sample properties from its spectroscopic readings. Because the spectroscopic readings at different wavelengths, especially at the adjacent wavelengths, are highly correlated, it has been shown that variable selection could significantly improve a soft sensor's prediction performance while reducing the model complexity. To improve the prediction performance, most variable selection methods focus on identifying the variables (i.e., wavelengths or wavelength segments) that are strongly correlated with the dependent variable. Although many successful applications have been reported, these variable selection methods do have their limitations. Specifically, the selected wavelengths sometimes show little connection to the chemical bounds or functional groups presenting in the sample. In addition, the selected variables can be quite sensitive to the choice of the training samples. In this work, we address these limitations from a different perspective: if a variable selection algorithm can identify the truly relevant input variables, it should consistently identify the same subset of variables regardless of the choice of the training samples. Therefore, we propose a variable selection method that aims to improve the consistency of variable selection resulted from different training samples. The new algorithm is termed consistency-enhanced evolution for variable selection (CEEVS). To demonstrate the performance and

robustness of CEEVS, we compare the proposed method with three representative variable selection methods using five published NIR datasets. These case studies clearly demonstrate that by improving the variable selection consistency, we can not only achieve improved prediction performance, but also identify key chemical information from spectroscopic data.

Key words: Variable selection; soft sensor; spectroscopic data; near-infrared

1 BACKGROUND

With the advancements of spectroscopic technologies including near-infrared (NIR), Ramon Spectroscopy, and UV/Vis spectroscopies, various properties could be inferred from a sample's spectrum profile. Correspondingly, multivariate modeling approaches (i.e., soft sensor models), which correlate the spectroscopic reading of a sample to its properties of interest, have drawn increased research interest. These soft sensor models offer a non-invasive, fast and cheap way to estimate the sample properties of interest and have been applied in many different fields. For example, spectra-based soft sensors have been developed to determine properties such as octane number of gasoline, moister content of corn, active pharmaceutical ingredient (API) in drug, and microorganism concentration in a mixed culture[1–4]. The most commonly used modeling approach for soft sensor is partial least squares (PLS) due to its simplicity, robustness and the inherent capability in addressing collinearity among independent variables.

It has been well-recognized that the performance of a soft sensor can be significantly improved if only the relevant variables are included as predictor[5–9]. This is particularly the case for spectrum-based soft sensors, where the readings at different wavelengths are highly correlated. In addition, most multivariate statistical methods, including PLS, require much larger number of samples than the number of variables to perform well. However, most spectral datasets have relatively small sample size (less than 100) but large number of variables (several hundreds of wavelength). Therefore, eliminating irrelevant wavelengths could help circumvent this difficulty by reducing the number of variables. Driven by these considerations, many variable selection methods have been developed in the past few decades. Most existing variable selection

methods focus on selecting the variables (i.e., wavelengths or wavelength segments) that are strongly correlated with the dependent variable to improve the prediction performance. These variable selection approaches include direct methods that rank variable contributions such as variable selection based on variable importance in projection (VIP)[10] or regression coefficient (BETA)[7], and iterative methods such as uninformative variable elimination (UVE)[11] and least absolute shrinkage and selection operator (Lasso)[12]. Among iterative approaches, a group of variable selection methods based on the principle of "survival of the fittest" have shown superior performance. The representative methods of this group are the genetic algorithm (GA)[13–15], the competitive adaptive reweighted sampling method (CARS)[3] and the method based on stability and variable permutation (SVP)[16]. By employing the principle of "survival of the fittest", these methods rely on random sampling in the variable space and/or sample space to identify the most relevant input variables as predictor variables to improve prediction performance.

Despite many successful applications, existing variable selection methods also have limitations. It has been recognized that a soft sensor model with good fitness performance may not guarantee good variable selection performance[5,7]. Specifically, for spectrum-based soft sensors, the selected wavelengths sometimes show little connection to the chemical bounds or functional groups presenting in the sample. In addition, the selected variables can be quite sensitive to the choice of the training and validation data. In particular, the variables selected from different Monte Carlo (MC) runs using randomly selected training and validation data often show low consistency with each other. The inconsistency among different MC runs suggests that the selected variables (wavelengths) may not contain the truly relevant predictors, i.e., the wavelengths corresponding to the underlying chemical bonds or functional groups that determine the property of the sample. To help address this limitation, in this work, we propose a new variable selection method from a different perspective: if a variable selection algorithm can identify the truly relevant input variables, it should consistently identify the same subset of variables regardless of the choice of the training dataset. The proposed method, namely consistency enhanced evolution for variable selection (CEEVS), aims to improve the consistency of variable selection across different training datasets. To examine whether

CEEVS is able to deliver improved prediction performance and to identify relevant chemical information, five case studies were presented. The performance of CEEVS is also compared with three representative methods based on the principle of "survival of the fittest", i.e., GA, CARS and SVP, using the full PLS model as the basis.

The rest of the paper is organized as follows. Section 2 briefly reviews the relevant existing methods; Section 3 describes the proposed CEEVS methods; Section 4 presents the results from 5 cases studies; and Section 5 is the conclusion and discussion.

2 REVIEW OF RELEVANT VARIABLE SELECTION ALGORITHMS

2.1 Genetic algorithm (GA)

Inspired by Darwin's evolution theory of "survival of the fittest", GA is one of the most commonly applied variable selection methods[14,15,17]. According to the evolution theory, the individuals who are well adapted to the environment will be more likely to survive and produce the next generation[13]. Therefore, in GA, parent chromosomes (i.e., subsets of selected variables) are determined based on its "fitness to the environment", such as prediction performance. Then crossover and mutation are applied to produce offspring, i.e., new sets of selected variables. Through crossover, portions of two parent chromosomes are crossed and combined to make two offspring which have new combinations of genes (i.e., variables or wavelengths); through mutation, new genes not included in the chromosomes population could have a chance to be included, which may improve the offspring's fitness to the environment. This reproduction step is repeated until a termination criterion is satisfied[18].

2.2 Competitive adaptive reweighted sampling (CARS)

In CARS, the importance of a variable is determined based on its absolute regression coefficient (BETA) obtained through partial least squares (PLS) regression. The variables with large absolute regression coefficients are considered as the important variables. CARS employs the iterative sampling runs to

determine the optimal subset of variables. In each sampling run, two variable reduction procedures, namely exponentially decreasing function (EDF) and adaptive reweighted sampling (ARS), are applied to reduce the number of variables. The root mean square error of cross-validation ($RMSE_{CV}$) is calculated using the variables retained in the sampling run. After n_S times sampling runs, CARS obtains n_S models consisting of the different subsets of variables and the model with the lowest $RMSE_{CV}$ is selected as the optimal model[3]. CARS has been applied to develop soft sensors in many different applications, including spectroscopic data collected from GC-MS, NIR, and UV/Vis[19–22].

2.3 Stability and variable permutation (SVP)

Recently, SVP was proposed based on the evolutionary principles of 'intraspecific competition' and 'survival of the fittest'. In SVP, the importance of each variable is determined through variable stability and variable permutation analysis. Variable stability is evaluated through random sampling of the sample space, while variable permutation analysis is performed through random sampling of the variable space. After computing the variable stability and performing variable permutation analysis, SVP divide all the variables into the elite variable set and normal variable set by adaptive reweighted sampling (ARS). The elite variable set consists of variables with high stability, while the normal set contains variables with relatively low stability. To eliminate the uninformative variables, SVP employs exponentially decreasing function (EDF), which remove variables with small difference from the normal variable set. In each sampling run, the procedures described above are performed. After n_S sampling runs, SVP obtains n_S models with different variable subsets; then the variable subset that results in minimum mean and relatively low standard deviation of the $RMSE_{CV}$'s is selected as the optimal subset of the selected variables[16].

3 THE PROPOSED METHOD: CONSISTENCY ENHANCED EVOLUTION FOR VARIABLE SELECTION (CEEVS)

As discussed earlier, the variables selected by GA, CARS and SVP are not necessarily the truly relevant variables, i.e., the ones corresponding to the key chemical bonds or functional groups that determine the

sample properties at interest. This is further reflected in the low consistency among variable selection results obtained from different MC runs using randomly selected training samples. To explore the root cause of this deficiency, we compare the evolution theory based variable selection to biological evolution. In biological evolution, it usually takes millions of years for natural selection to converge to an optimal solution; however, in variable selection the limited sample space and the limited evolution process may cause the variable selection to be stuck in a local optimum and miss the global one. Therefore, we believe that the limited sample space and limited evolution may be one of the underlying reasons for the inconsistency among different MC runs. However, without knowing what the global optimum is (i.e., the ground truth of the truly relevant variables), it is difficult to devise approaches to directly address this limitation.

In this work, we address this difficulty based on the following rationale: if a variable selection algorithm can identify the truly relevant input variables, it should consistently identify the same subset of the variables regardless of the choice of the training samples. In other words, the variable selection results among different MC runs should be relatively consistent to identify the truly relevant predictors. Therefore, we hypothesize that if a variable selection method delivers better consistency in terms of selected variables among different MC runs, it is more likely that it selects the truly relevant variables and as a result would deliver better prediction performance. Based on this hypothesis, we propose the CEEVS algorithm aiming to improve the consistency in variable selection. Below we first introduce the necessary notations, then present the details of the algorithm.

3.1 Notation

Spectral data and PLS model: In this work, $\boldsymbol{X}_{n\times m}$ denotes the spectral data, which consists of n samples and spectral absorbances of m wavelengths for each sample; $\boldsymbol{Y}_{n\mathrm{x}l}$ denotes the l properties of interest for the n samples. Both $\boldsymbol{X}_{n\times m}$ and $\boldsymbol{Y}_{n\mathrm{x}l}$ are autoscaled to zero mean and unit variance before model development through PLS. In the PLS model, the regression equations are the following

$$\boldsymbol{X}_{n\times m} = \boldsymbol{T}_{n\times p}\boldsymbol{P}^{\mathrm{T}}_{m\times p} + \boldsymbol{E}_{n\times m} \quad (1)$$

$$\boldsymbol{Y}_{n\times l} = \boldsymbol{U}_{n\times p}\boldsymbol{Q}^{\mathrm{T}}_{l\times p} + \boldsymbol{F}_{n\times l} \quad (1)$$

where p is the number of principal components; $\boldsymbol{T}_{n\times p}$ and $\boldsymbol{U}_{n\times p}$ are the score matrices; $\boldsymbol{P}_{m\times p}$ and $\boldsymbol{Q}_{l\times p}$ are the loading matrices; $\boldsymbol{E}_{n\times m}$ and $\boldsymbol{F}_{n\times l}$ are the error or residual matrices, respectively. The PLS model maximizes the covariance between $\boldsymbol{T}$ and $\boldsymbol{U}$.

Gene, chromosome and fitness: In CEEVS, many notations follow the GA method. A gene refers to an individual variable (wavelength), and a chromosome ($C_{m\times 1}$) refers to a set of selected variables: the *i*-th element (c_i) of the chromosome is either "1" or "0", indicating whether the *i*-th variable is included in the chromosome or not, respectively. The fitness of a chromosome is determined through prediction error, i.e., normalized root mean squared error from cross-validation ($NRMSE_{CV}$).

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

where, n_V is the number of samples of the validation dataset. In this work, 10-fold cross validation is employed for all methods. Therefore, the average of the ten $NRMSE_{CV}$'s is used.

Variable stability: In existing literature[3,16,23], variable stability is determined through random sampling of the training data and evaluating how consistently the variable contributes to the soft sensor model. Specifically, to compute the stability, MC sampling is applied in which certain percentage (denoted as γ) of the n samples are randomly selected to build a PLS model, and this random selection is iterated for n_S times. A full PLS model that include all wavelengths as input variables is established for each subset of data to compute regression coefficients. As regression coefficient (BETA) determines how much a variable contribute to the prediction of the dependent variable, it has been used to evaluate the stability of each variable, as shown in Eqn. (3).

$$S_{BETA-j} = \frac{|\bar{b}_j|}{\sqrt{\frac{1}{n_S-1}\sum_{i=1}^{n_S}(b_{ij}-\bar{b}_j)^2}} \quad (3)$$

where, S_{BETA-j} is the stability of the j-th variable based on regression coefficients, $\bar{b}_j$ is the average value of regression coefficients of j-th variable from n_S full PLS models using samples randomly selected from the training dataset based on the pre-determined sampling ratio γ, and b_{ij} is the regression coefficient of j-th variable in i-th PLS model.

Besides regression coefficient BETA, variable importance in projection (VIP) also indicates how much a variable contributes to the dependent variable. Unlike BETA, VIP scores estimate the importance of each variable in the projection used in a PLS model. It has been reported that when each predictor contributes differently to the dependent variable (which is the case for most, if not all, practical applications), BETA-based variable selection may not work as well as VIP-based variable selection[5,7]. In fact, Wold et al.[10] recommended a combination of VIP and BETA for variable selection. To improve the consistency of variable selection, in this work we propose using the combination of VIP and BETA to compute variable stability. To do so, we first define variable stability based on VIP, as shown in Eqn. (4).

$$S_{VIP-j} = \frac{|\bar{v}_j|}{\sqrt{\frac{1}{n_S-1}\sum_{i=1}^{n_S}(v_{ij}-\bar{v}_j)^2}} \quad (4)$$

where S_{VIP-j} is the stability of the j-th variable based on VIP scores, $\overline{v_j}$ is the average value of the VIP scores of j-th variable among n_S models, and v_{ij} is the VIP score of j-th variable in the i-th model. To combine S_{BETA} and S_{VIP} for determining the stability for each variable, S_{BETA} and S_{VIP} are first standardized since they have different scales.

$$Z_{BETA-j} = \frac{S_{BETA-j} - \overline{S_{BETA}}}{std(S_{BETA})} \quad (5)$$

$$Z_{VIP-j} = \frac{S_{VIP-j} - \overline{S_{VIP}}}{std(S_{VIP})} \quad (6)$$

where $\overline{S_{BETA}}$ and $\overline{S_{VIP}}$ are the average stability of all variables based on BETA and VIP scores, respectively; $std(S_{BETA})$ and $std(S_{BETA})$ are the corresponding standard deviations, respectively. Then the average of Z_{BETA-j} and Z_{VIP-j}, denoted as Z_j, is used to determine the stability of the j-th variable.

$$Z_j = \frac{1}{2}(Z_{BETA-j} + Z_{VIP-j}) \tag{7}$$

Note that in this work, Z_{BETA} and Z_{VIP} were assigned the same weight, which can be adjusted for different applications.

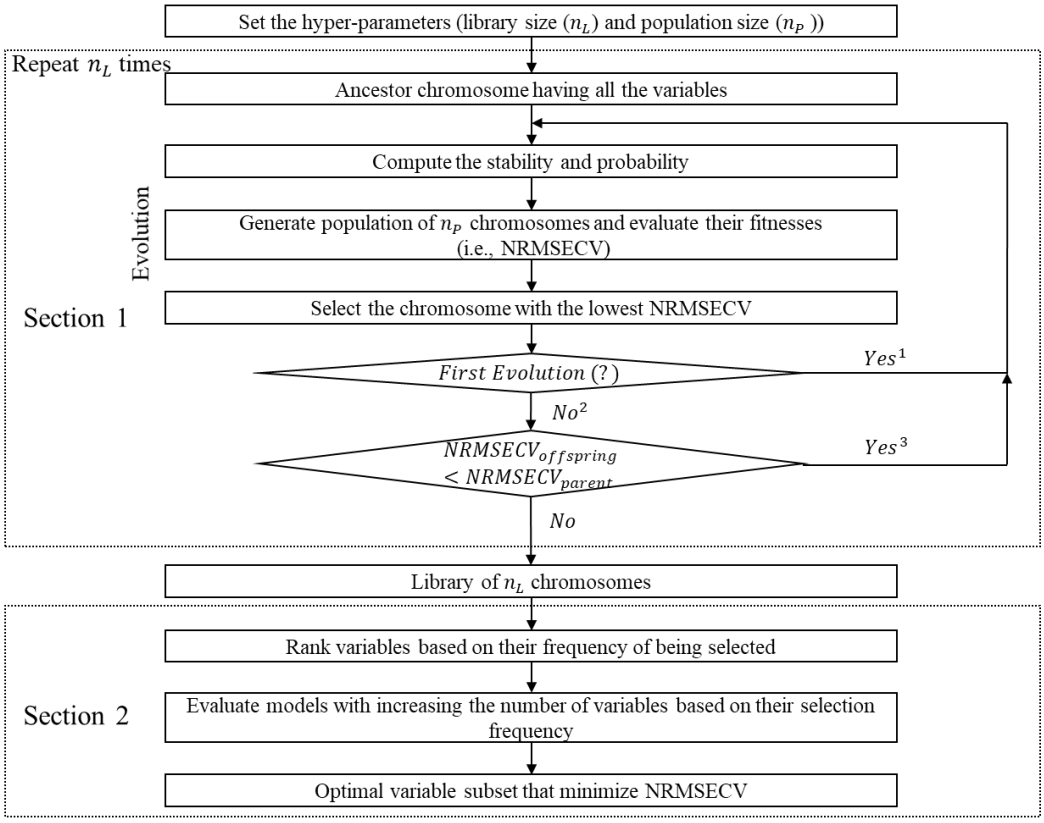
Probability of selection: in CARS and SVP, the variable stability is used to directly determine how often a variable is included in the initial population. In this work, to remove any potential bias, we first convert the variable stability into a probability; then each variable is randomly selected according to its probability to generate the initial population of chromosomes. The probability of the j-th variable is defined as followings:

$$p_j = \lambda_1 + (\lambda_2 - \lambda_1)\left(\frac{Z_j - Z_{min}}{Z_{max} - Z_{min}}\right) \tag{8}$$

where, λ_1 is a small probability (10^{-5} in this work) to ensure that even the variable of the minimum stability has a chance to be selected and evaluated; λ_2 is 1; Z_{max} and Z_{min} are the maximum and minimum stabilities among all variables, and Z_j is the stability of the j-th variable. When $Z_j = Z_{min}$, $p_j = \lambda_1$; when $Z_j = Z_{max}$, $p_j = \lambda_2 = 1$.

3.2 Outline of the CEEVS

As shown in Fig. 1, CEEVS consists of two main sections: Section I is to construct a library with optimal chromosomes, and Section II is to select the optimal subset of variables from the library to build the soft sensor.



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. 1. Flow diagram of the CEEVS algorithm

For Section I, CEEVS takes a consistency enhanced evolution process in order to obtain an optimal chromosome with limited iterations. In GA, the chromosomes of the initial population are generated randomly where each variable has the same probability to be selected. In CEEVS, starting with the complete variable set, the initial chromosome population is generated randomly based on each variable's probability of selection as defined in Eqn. (8). As shown in the previous section, the probability of selection is simply a scaled variable stability; in other words, variables with higher stability will be selected with higher probability. 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. Once the initial population of n_P chromosomes are obtained, each chromosome is evaluated for its fitness value. In this work, we use the

selected variables (i.e., the variables that have "1" in the chromosome) to build a PLS model, and the model's $NRMSE_{CV}$ value is used as the fitness value for the chromosome. The optimal chromosome, i.e., the one with the minimal $NRMSE_{CV}$ within the initial population, is considered 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. Again, 10-fold cross validation is employed in this work for all methods. Therefore, $NRMSE_{CV}$ is actually referring to the average of the ten $NRMSE_{CV}$'s.

The evolution process of CEEVS is completely different from GA. Instead of cross-over and mutate, in CEEVS, we simply use the variables selected by the parent chromosome as the new complete 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 considered the parent chromosome, and the variable selected by the parent chromosome is considered as the new "full" variable set; Next, the variable stability and probability are re-computed for this new "full" set; then, a population of n_P offspring are generated randomly based on the variable's probability for selection, and evaluated for their fitness value. In this way, all the offspring are guaranteed to contain fewer variables than the parent and may have a better fitness value. This evolution process is repeated until the fitness of the offspring is worse than that of the parent, meaning the parent can no longer produce better offspring. Then the parent of the final evolution run, i.e., the best chromosome generated from the evolution process, is stored into the library. This evolution process will repeat n_L times with different random seeds, which is the pre-determined library size, i.e., the number of the optimal chromosomes to be stored in the library. Each time the process starts with the complete set of variables. At the end of n_L repetitions, the library will contain n_L optimally evolved chromosomes, i.e., subsets of selected variables that deliver the lowest $NRMSE_{CV}$ during each repeated evolution process.

For Section II, starting with the library that contains n_L best chromosomes generated in Section I, 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 the 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. This upper limit can be adjusted to reduce the risk of overfitting. In this work, we set the upper limit as 300 variables. Finally, all models are evaluated for their fitness ($NRMSE_{CV}$), and the variable subset that produce lowest $NRMSE_{CV}$ value is considered the final result of the selected variables.

It is worth noting that all Monte Carlo (MC) repetitions involved in the CEEVS and other variable selection methods are carried out on the training samples only. Specifically, the procedures of CEEVS shown in Fig. 1 were all performed using the training samples only, with n_L MC repetitions of different random seeds to generate library of n_L chromosomes.

3.3 The choice of tuning parameters

One of the advantages of CEEVS is simpler tuning compared to GA. First, there are only four parameters in CEEVS, which include the library size (n_L), the population size (n_P), the ratio of samples (γ) and the number of sampling runs (n_S). n_L determines the number of the chromosomes to be stored in the library, which is also the number of repetition (or evolution) in Section I of the algorithm. n_P is the number of chromosomes present in each population. γ and n_S are related to evaluating variable stability: γ is the ratio or percentage of samples to be randomly selected and n_S is the number of the randomly selected sample subsets, i.e., the number of PLS models to be built in order to evaluate the variable stability. Second, CEEVS is not sensitive to these parameters. As detailed later in Sec. 5.3, sensitivity analysis show that when n_L, n_P, γ and n_S are large enough, their effect on the final soft sensor performance becomes negligible. Therefore, in this work we decide to keep all 4 parameters fixed instead of changing them from dataset to dataset. Table 1 lists the parameter setting used in this work, and the recommended range if one chooses to fine tune the parameter.

Table 1. Parameters used in this work and recommended range of tuning parameters

	Parameter (this work)	Recommended range
n_L	200	100 – 500
n_P	400	200 – 500
γ	0.9	0.8 – 0.9
n_S	400	300 – 800

4 CASE STUDIES

In this section, we use five near-infrared (NIR) datasets to demonstrate the performance of the proposed CEEVS method, which is compared with three representative "survival of the fittest" based methods: CARS, SVP, and GA. The full PLS model that utilizes all the variables in the NIR spectrum is used as the comparison basis.

4.1 Datasets

Five published NIR datasets are used to evaluate the performance of different variable selection methods. Table 2 summarizes the five datasets, including the number of samples and variables, the partition of the dataset into training and testing, as well as relevant references. Figure 2 plots the sample spectra for each dataset.

Table 2. Summary of the five NIR datasets

	# of samples in calibration set	# of samples in test set	# of samples in total	# of variables	Property of interest	Reference
Corn[a]	64 (80%)	16 (20%)	80	700	Protein content	[2,24]
Diesel	180 (70%)	76 (30%)	256	401	Aromatic content	[16,25]
Pharma	459 (70%)	196 (30%)	655	650	Active pharmaceutical ingredients (API)	[2,4,26]
Wheat	121 (80%)	30 (20%)	151	150	Protein concentration	[16,27]
Beer	48 (80%)	12 (20%)	60	926	Extract concentration	[28,29]

[a]NIR spectra measured on mp5 spectrometer was used.

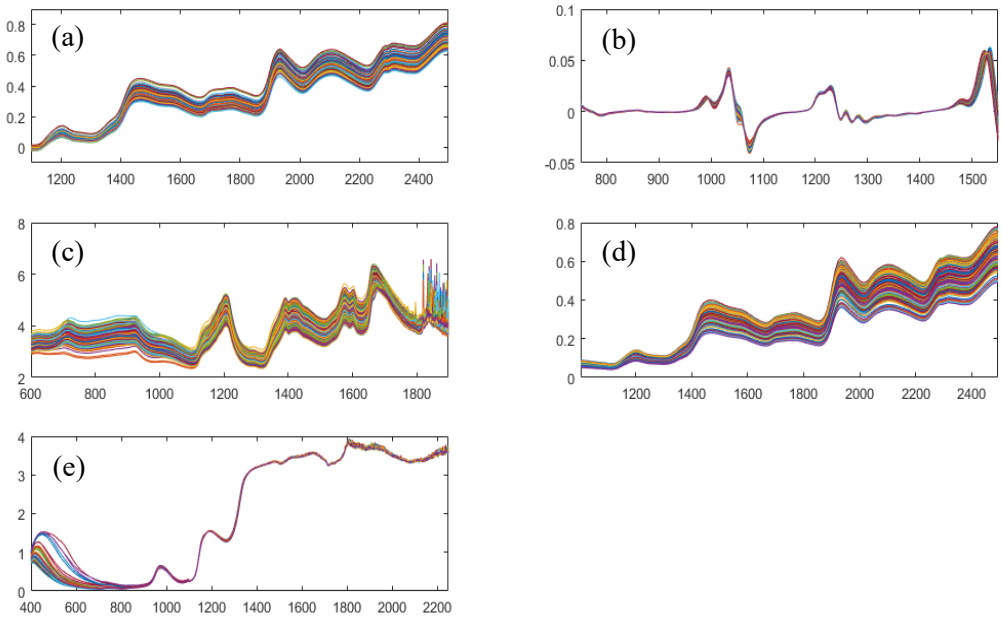


Fig. 2. The spectra of five datasets. (a) corn dataset; (b) diesel fuel dataset; (c) pharmaceutical tablets dataset; (d) wheat dataset; (e) beer dataset. For all subplots, x-axis is wavelength (nm) and y-axis is absorbance.

4.2 Simulation setup and performance metrics

To eliminate the potential bias caused by a specific partition of the whole dataset into calibration and testing subsets, a Monte Carlo validation and testing (MCVT) procedure that we proposed previously is followed[2]. Specifically, 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 2.

The performance of different variable selection methods is assessed through three metrics. The first two are based on the soft sensor prediction performance, while the third directly evaluates the performance of variable selection through a consistency index.

We choose normalized root mean square error in prediction ($NRMSE_P$) to evaluate the prediction performance of different soft sensor models. The definition of $NRMSE_P$ is given in Eqn. (9), where, n_T is

the number of samples of the test dataset in each MC runs. As shown in Eqn. (9), the normalization in $NRMSE_P$ facilitates the comparison of different methods across different datasets.

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

In this work, the mean and the standard deviation of $NRMSE_P$ obtained from the 100 MC runs are used as the two metrics to evaluate the performance of different methods. The mean ($\overline{NRMSE_P}$) evaluates the accuracy of each method while the standard deviation (σ_{NRMSE_P}) assesses the robustness of the method[2].

To evaluate the consistency of the variable selection among different MC runs, we define a consistency index (I_c) as the following:

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

where m' is the number of the variables (among all m variables) being selected at least once among all MC runs; $prob(x_i)$ is the probability of the i-th variable being selected, which is quantified by how frequently a variable is selected among all the MC runs. Clearly, a higher I_c indicates that the informative variables are more consistently selected regardless of calibration datasets.

It is worth noting that if a final model is built using the summative/cumulative results of the 100 MC runs, e.g., variables selected/included in the model and/or number of principal components used by the PLS model, there will be an issue of "using the testing data as the training data" in an ad-hoc way, which could lead to overfitting. However, this is not the case in this work. Each MC run is independently conducted and evaluated, with clear separation of training and testing samples. The outputs of each MC run are $NRMSE_P$, variables selected, number of PC's used in the PLS model. These results of one MC run have no influence on the results of other MC runs. In the end, the average and standard deviation of all $NRMSE_P$'s from 100 MC runs, as well as I_c are used to evaluate the performance of different variable selection methods. These summative results have no influence on the process of variable selection or PLS model building.

It is also worth noting that different MC runs will result in different variables being selected due to different training samples being used and the stochastic nature of all "survival of the fittest" based variable selection methods. When these selected variables are used to build PLS models, the principal components (PC's) will be different for different MC runs. It is also possible that the number of PC's will be different as it is determined through 10-fold cross-validation. The goal of MCVT is to compare different variable selection methods through the accuracy (i.e., the average of the 100 $NRMSE_P$'s) and precision or robustness (i.e., the standard deviation of the 100 $NRMSE_P$'s) of each method. A similar approach has been reported in the literature [30].

5 RESULTS AND DISCUSSION

To ensure a fair comparison, all methods being compared were optimized through 10-fold cross-validation. The tuning parameters for each method are listed in Table 3. For each method, the optimal tuning parameters were determined through exhaustive search within a specified range for the parameter.

Table 3. Tuning parameters that were optimized for each method

Methods	**Tuning parameters**
Full PLS	# of PC's
CARS	# of PC's, # of Monte Carlo sampling runs
SVP	# of PC's, # of iterations, sampling ratio of MCS-S[a] and MCS-P[b], # of sampling in MCS-S[a] and MCS-P[b]
GA	# of PC's, population size, # of iterations, crossover scheme, mutation rate, initial population, termination criterion
CEEVS[c]	# of PC's

[a]Monte Carlo sampling in sample space; [b]Monte Carlo sampling in variable space; [c]Other parameters are fixed as shown in Table 1.

5.1 Performance comparison

For each dataset, the variable selection and soft sensor prediction results from each method are tabulated in Table 4 – 8. The best performance corresponding to each metric is shown in boldface. In these tables, Improvement rate (%) refers to the improvement of $\overline{NRMSE_P}$ over that of the full PLS model, n_{PC} is the "mean ± std" of the number of principal components of the final soft sensor among 100 MC runs, n_{VAR} is the "mean ± std" of the number of selected variables among 100 MC runs, except full PLS where all variables are used.

Table 4. The performance comparison using the corn dataset.

Method	$\overline{NRMSE_P}$	σ_{NRMSE_P}	I_C	**Improvement rate (%)**	n_{PC}	n_{VAR}
Full PLS	9.197	2.390	-	-	11.6 ± 1.7	700
CARS	9.263	2.760	0.063	-0.72	12.3 ± 1.7	21.4 ± 8.2
SVP	9.569	2.602	0.062	-4.05	14.0 ± 0.9	25.9 ± 10.0
GA	8.730	2.337	0.119	5.07	9.0 ± 2.4	73.6 ± 27.2
CEEVS	**8.335**	**2.051**	**0.212**	9.37	9.1 ± 2.3	100.9 ± 39.2

Table 5. The performance comparison using the diesel fuel dataset.

Method	$\overline{NRMSE_P}$	σ_{NRMSE_P}	I_C	**Improvement rate (%)**	n_{PC}	n_{VAR}
Full PLS	2.38	**0.30**	-	-	12.3 ± 1.7	401
CARS	2.94	0.65	0.136	-23.54	13.1 ± 1.6	54.7 ± 55.1
SVP	2.32	0.43	0.150	2.71	13.6 ± 1.4	47.0 ± 13.9
GA	2.24	**0.30**	0.240	6.12	11.8 ± 1.7	92.0 ± 41.5
CEEVS	**2.20**	**0.30**	**0.432**	7.56	11.2 ± 1.8	123.4 ± 37.5

Table 6. The performance comparison using the pharmaceutical tablets dataset.

Method	$\overline{NRMSE_P}$	σ_{NRMSE_P}	I_C	Improvement rate (%)	n_{PC}	n_{VAR}
Full PLS	5.05	**0.76**	-	-	14.3 ± 2.5	650
CARS	4.72	0.84	0.064	6.50	15.1 ± 3.1	30.2 ± 15.0
SVP	4.85	0.83	0.104	3.85	18.5 ± 1.5	50.1 ± 25.8
GA	4.46	0.90	0.138	11.69	10.8 ± 3.0	69.1 ± 44.1
CEEVS	**4.45**	0.89	**0.231**	11.86	13.3 ± 2.4	91.9 ± 56.1

Table 7. The performance comparison using the wheat dataset.

Method	$\overline{NRMSE_P}$	σ_{NRMSE_P}	I_C	Improvement rate (%)	n_{PC}	n_{VAR}
Full PLS	3.614	**0.587**	-	-	15.9 ± 1.5	150
CARS	3.687	0.669	0.243	-2.02	15.2 ± 2.1	36.3 ± 13.0
SVP	4.011	0.685	0.151	-11.00	18.0 ± 1.7	21.8 ± 2.5
GA	3.502	0.595	0.286	3.08	10.7 ± 1.7	40.4 ± 13.6
CEEVS	**3.497**	0.624	**0.289**	3.22	11.2 ± 2.4	35.5 ± 11.4

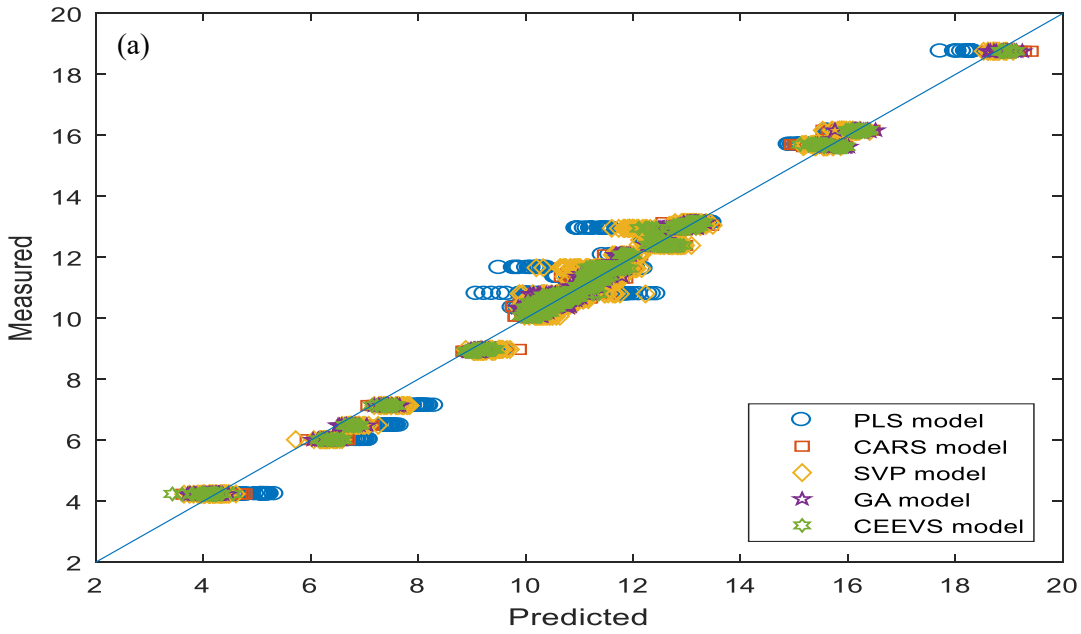
Table 8. The performance comparison using the beer dataset.

Method	$\overline{NRMSE_P}$	σ_{NRMSE_P}	I_C	Improvement rate (%)	n_{PC}	n_{VAR}
Full PLS	6.57	6.46	-	-	9.1 ± 2.6	926
CARS	3.24	2.76	**0.192**	50.64	9.1 ± 2.6	86.8 ± 38.2
SVP	4.18	5.20	0.166	36.28	13.4 ± 2.1	113.0 ± 12.6
GA	2.37	1.85	0.142	63.91	7.8 ± 2.6	94.1 ± 58.0
CEEVS	**2.36**	**1.45**	0.182	64.11	8.1 ± 2.6	130.2 ± 85.9

As shown in the tables, across different datasets, CEEVS performs the best in almost all performance metrics. Specifically, among all 15 comparison instances (5 dataset × 3 performance metrics). In terms of

$\overline{NRMSE_P}$, CEEVS performs the best for all 5 datasets; in terms of I_C, CEEVS performs the best for 4 of the 5 datasets and the 2nd best for the rest one; in terms σ_{NRMSE_P}, CEEVS performs the best for 3 of the 5 datasets, while slightly larger σ_{NRMSE_P} for the rest 2 datasets. These results indicate that by enhancing the consistency of variable selection, we can achieve better prediction performance.

Besides the quantitative metrics given in the tables, Figure 3 (a) and (b) compare the predicted vs measured quality variable for the diesel and beer datasets. From these two figures, it can be seen that the predictions of CEEVS stay the closest to the diagonal line, further indicating the superior prediction accuracy and robustness.



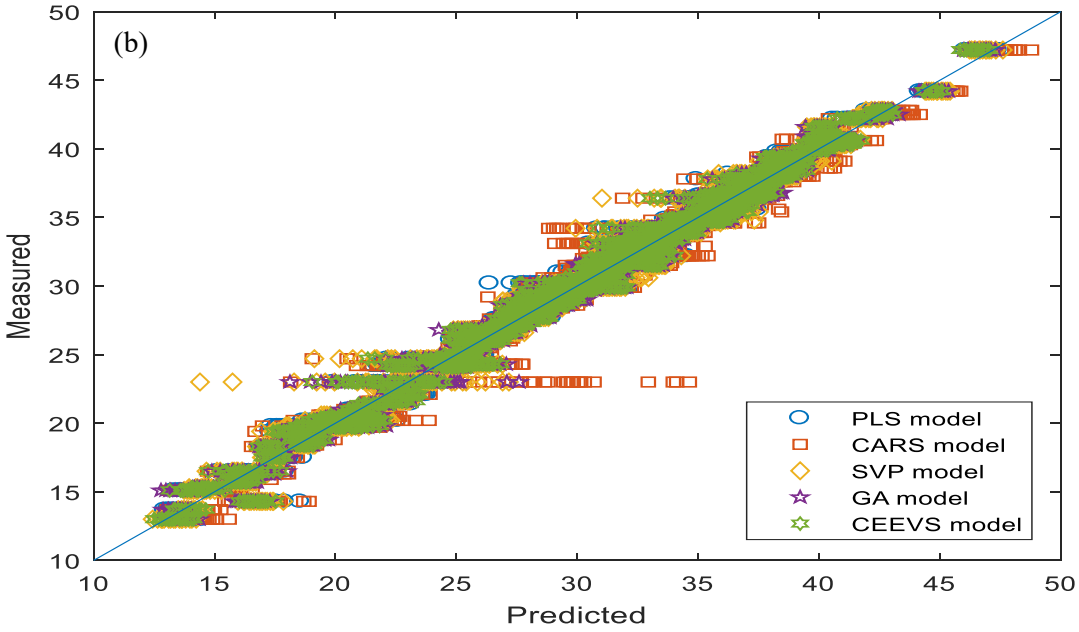


Fig. 3. Plot of predicted vs. measured properties from five methods. (a) beer dataset; (b) diesel dataset.

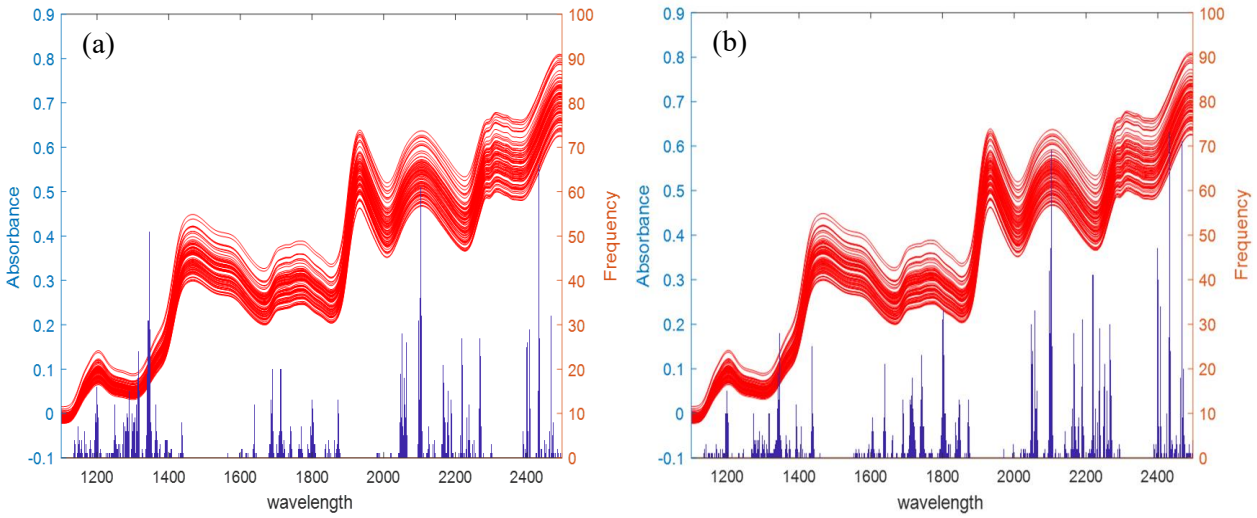
5.2 CEEVS can extract the underlying chemical information

As discussed in Sec. 1, one of the limitations of the existing variable selection methods is that the selected variables (wavelengths) for the soft sensor model may not have clear relationship with the chemical bounds or functional groups presenting in the sample. By enhancing the consistency of variable selection, we expect that CEEVS could identify the truly relevant variables that reveal the underlying chemical information. Further examination of the variable selection results from different methods confirmed our hypothesis. Due to limited space, here we use results from two dataset to illustrate this in detail and provide the results for the other datasets in the Supporting Information.

Figure 4 and 5 plot the frequency of each variable being selected (denoted by the vertical thin bars) among all 100 MC runs for the corn dataset and the pharmaceutical tablets dataset for all four variable selection methods. The sample spectra (denoted by the red curves) are plotted on the same figures to visualize the

portions of the spectra that are selected at high frequencies by different variable selection methods. These figures clearly show that CEEVS delivers the best consistency in terms of variable selection, as the variables that were selected from different runs are clustered together around spectrum peaks/valleys at high frequency, indicating high consistency. More importantly, further analysis show that the selected variables (corresponding to peaks or valleys) are associated with different chemical bonds/groups, which are labelled on the plot for the CEEVS method. The underlying chemical information revealed by the selected variables further support our claim that the selected variables with high consistency are likely the truly relevant ones.

In terms of variable selection frequency, GA performs similar to CEEVS, while the clustering of the selected variables may not be as clear and distinct as that from CEEVS. For CARS and SVP, although the number of variables being selected by these two methods are usually much smaller than those from GA and CEEVS, the consistency of variable selection is much worse and as a result, the selected variables could reveal little underlying chemical information.



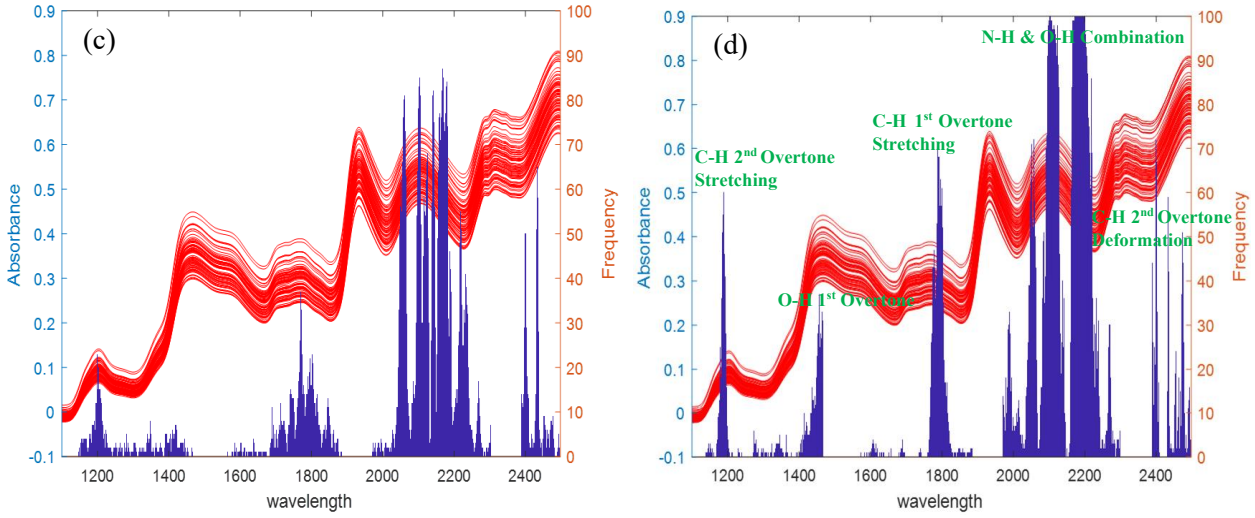
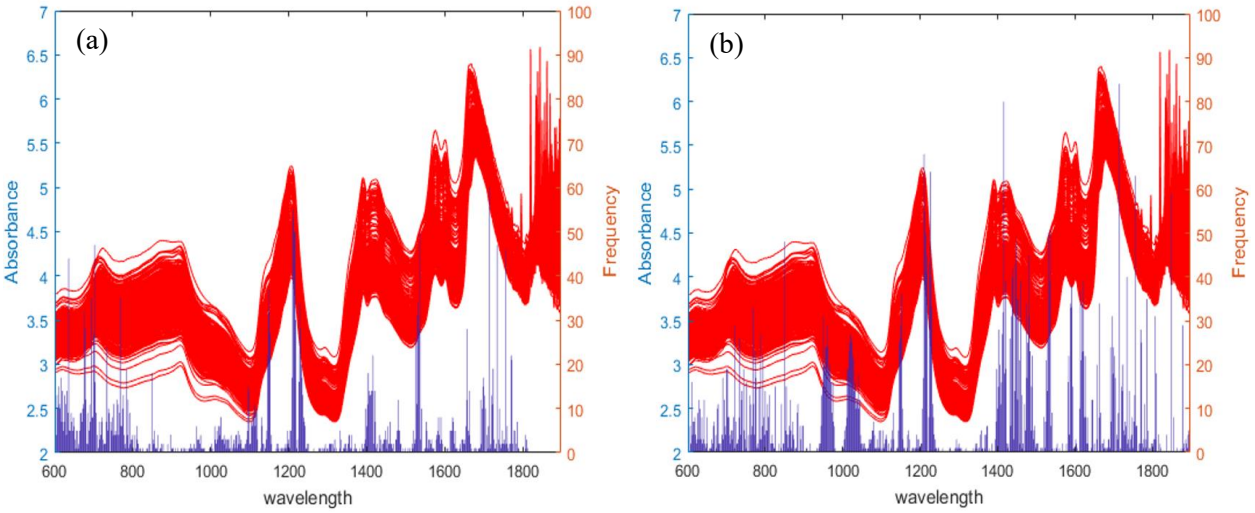


Fig. 4. Plot of spectra (red curves) and histogram of selected wavelengths (blue vertical bars) over 100 MC runs for the corn dataset. (a) CARS; (b) SVP; (c) GA; (d) CEEVS



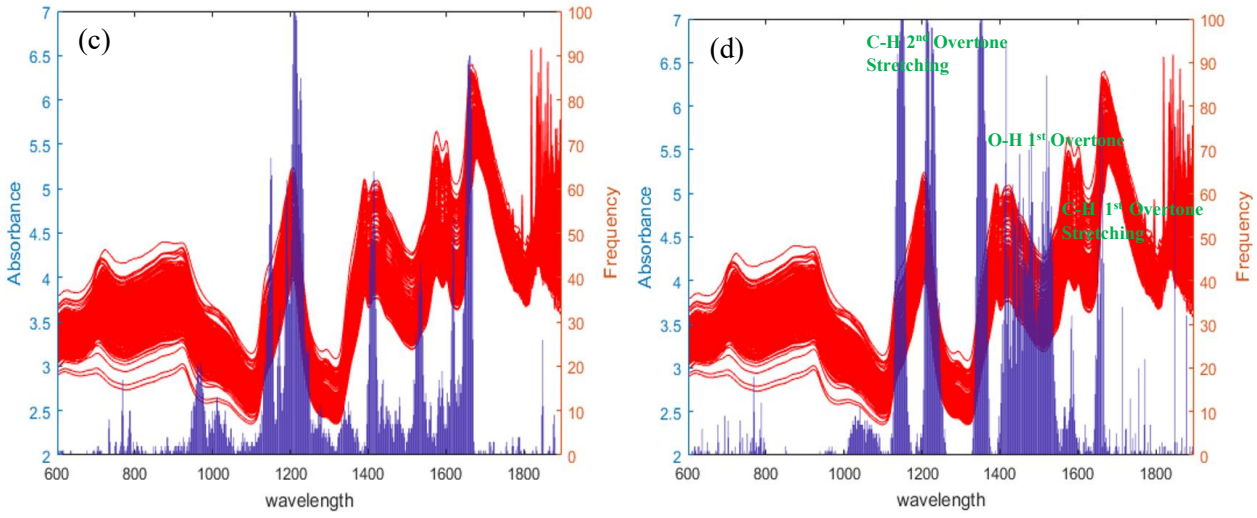


Fig. 5. Plot of spectra (red curves) and histogram of selected wavelengths (blue vertical bars) over 100 MC runs for the pharmaceutical tablets dataset. (a) CARS; (b) SVP; (c) GA; (d) CEEVS

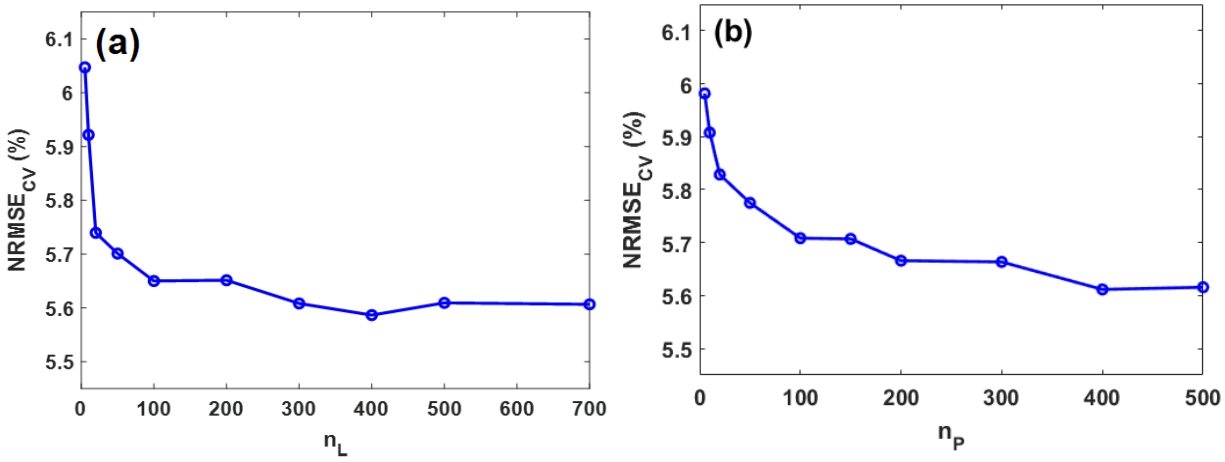
5.3 Robustness of CEEVS

CEEVS has four tuning parameters, the library size (n_L), the population size (n_P), the sampling ratio (γ) and the number of sampling runs (n_S). To examine the robustness of the method with respect to its tuning parameters, in this section, we test 10 different levels for each tuning parameter.

For the number of chromosomes in the library (n_L), the ten levels we tested were [5, 10, 20, 50, 100, 200, 300, 400, 500, 700]. The cross-validation results corresponding to the tested levels for the corn dataset is plotted in Figure 6 (a). The results for other datasets are very similar to the corn dataset, therefore they are omitted here. Figure 6 (a) shows that as n_L increase, $NRMSE_{CV}$ initially decreases sharply; and then it stabilizes when n_L is sufficiently large. Because n_L determines the number of best performing chromosomes to be stored in the library, the initial increase in n_L allows more relevant variables to be stored in the library; however, as n_L increasing, the enhanced variable selection consistency delivered by CEEVS allows all truly relevant variables being selected, therefore, further increasing the number of repetitions does not result in further

improvement in the model performance. Based on the testing of all datasets, in this work, we fix n_L at 200 for all the case studies.

For the size of population (n_P), the ten levels we tested were [5, 10, 20, 50, 100, 150, 200, 300, 400, 500]. The cross-validation results for the corn dataset is plotted in Figure 6 (b) and other dataset show very similar behavior. Similar to the case of n_L, as n_P increases, the cross-validation performance saw significant improvement initially, then levels off as n_P keep increasing. This is because the initial increase in n_P allows more chromosomes to be evaluated, thereby increasing the probability of producing superior offspring. However, after sufficient number of chromosomes have been evaluated, this effect diminishes. Based on the effect of n_P for all the datasets, we set n_P to 400 for all the case studies in this work.



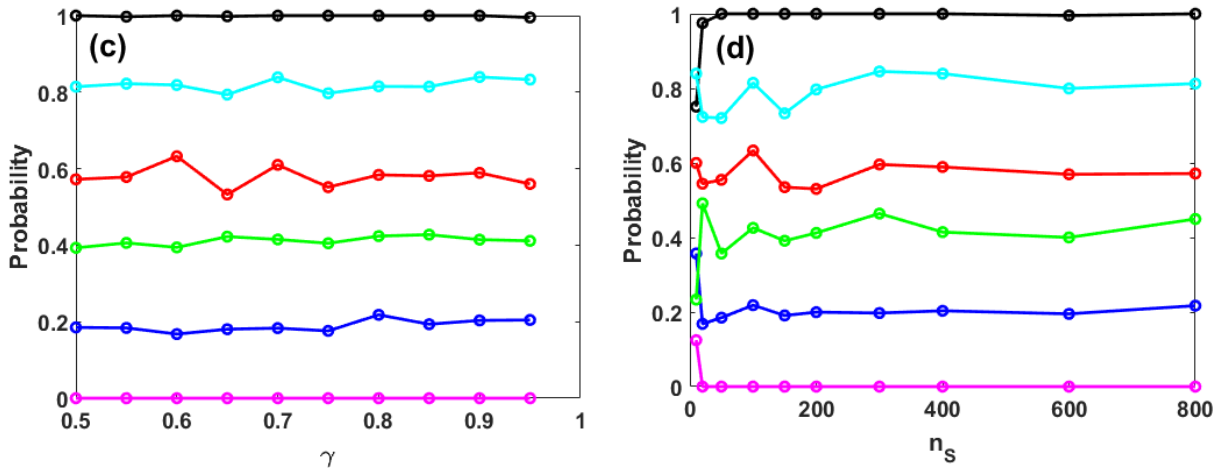


Fig. 6. (a) The effect of n_L on performance for the corn dataset. (b) The effect of n_P on performance for the corn dataset. (c) The effect of γ on the initial selection probability of five representative variables (denoted by different lines) that have different levels of probability of selection. (d) The effect of n_S the initial selection probability of five representative variables (denoted by different lines) that have different levels of probability of selection.

The sampling ratio (γ) and the number of sampling runs (n_S) are involved in evaluating variable stability and probability for selection, so here we examine their effect on variable's probability for selection. We selected 5 representative variables that have different levels of probability for selectin. For γ, the 10 levels examined are [0.50, 0.55, 0.60, 0.65, 0.70, 0.75, 0.80, 0.85, 0.90, 0.95], and for n_S, the 10 levels examined are [10, 25, 50, 100, 150, 200, 300, 400, 600, 800]. As shown in Figure 6 (c) and (d), similar to n_L and n_P, when γ and n_S are large enough, the probability for selection become quite insensitive to the tuning parameters. In this work, we choose γ = 0.9 and n_S = 400 for all case studies.

5.4 Discussion

It has been well documented that variable selection can help address several challenges associated with soft sensor development for spectroscopic datasets, namely: (1) variable multicollinearity, i.e., variables are highly correlated; (2) highly noisy data; (3) curse of dimensionality, i.e., the number of variables is larger

than the number of samples. In addition, variable selection could improve model predictive accuracy by eliminating irrelevant input variables and provide a better understanding of the chemically important wavelength regions by reducing model complexity. However, variable selection methods can be sensitive to calibration data and their performance may be unstable. As shown in Tables 4 – 8, PLS soft sensors using variables selected by CARS and SVP delivered worse prediction performance compared to the full PLS soft sensor without variable selection for 3 out of the 5 datasets. More importantly, the low consistency of selected variables among different MC runs suggests that their performances are sensitive to the choice of the training samples. There are two possible reasons to explain such sensitivity. First, both CARS and SVP use the regression coefficients to define the stability of variables, which introduces significant variability in variable selection as regression coefficients are sensitive to the choice of the training samples. Second, both methods adopt EDF to remove the less important variables. Once the variables are eliminated based on their stability (which depends heavily on the training samples), they will not be re-evaluated. However, some previously eliminated variables could contribute significantly to prediction when variable combination changes.

To address these limitations, in CEEVS both regression coefficients and VIP scores are used to define the variable stability; and by using the frequency of a variable being stored in the library to rank the variables instead of using variable stability, CEEVS allows less important variables to be evaluated in different combinations. In addition, unlike GA where the initial population is generated completely randomly, CEEVS uses variable stability to guide the generation of the initial population which favors the more important variables. Moreover, the evolution process in CEEVS is also guided by variable stability, which enables CEEVS to deliver much enhanced consistency in variable selection. We believe such enhanced consistency in variable selection suggests truly relevant variables are selected, as the underlying relationship between sample spectrum and sample property does not change across different training samples. As expected, the enhanced consistency in variable selection not only resulted in the improved soft sensor prediction performance, but also revealed key chemical information in the spectra. Finally, compared

to GA, CEEVS significantly reduces the number of tuning parameters and deliver highly robust performance over a wide range of turning parameters. This is highly desirable as it makes the implementation of CEEVS significantly easier for practitioners and could be adopted easily for different applications.

It is worth noting that because CEEVS evaluates fitness ($NRMSE_{CV}$) for each round of evolution, CEEVS takes longer to execute. As computation time increase linearly with n_L, we chose smaller n_L (200) in this work, which was sufficient to ensure CEEVS's superior performance in all case studies. Since variable selection is run off-line, we do not think computation would limit the application of CEEVS.

6 CONCLUSION

In the last few decades, many spectral-based soft sensors have been developed to predict sample properties from its spectroscopic reading. As spectroscopic readings from different wavelengths, especially from adjacent wavelengths, are often highly correlated, variable selection could significantly improve soft sensor prediction performance while reducing model complexity. This work presents a new variable selection method, namely consistency-enhanced evolution for variable selection (CEEVS). Similar to GA, CARS and SVP, CEEVS employs Darwin's evolution theory of "survival of the fittest" to select the relevant variables as predictors for the model. However, CEEVS is different from the other methods in the sense that CEEVS aims to improve the consistency of variable selection from different, randomly-selected training datasets. We hypothesize that if a variable selection method delivers better consistency in selected variable across different training samples, it would deliver better prediction performance. This is because the truly relevant variables will not change as a result of different training datasets. Therefore, if a variable selection algorithm can identify the truly relevant variables, it should consistently identify the same subset of variables regardless of the choice of the training dataset.

To enhance the consistency of variable selection, CEEVS uses both PLS regression coefficients (BETA) and variable importance in projection (VIP) to determine variable stability, which reduces the sensitivity to

the training data while the probability of selection based on the variable stability ensures that even the variable of the minimum stability has a chance to be selected. The probability of selection based on the variable stability also ensures that the evolution process will start with a better initial population than GA where the initial population is completely randomly selected. This helps the evolution to converge to the optimal faster. In addition, the chromosome evolution process is also different from GA. By using the parent chromosome from previous evolution run as the new starting point to re-evaluate the variable stability, and using the updated stability to determine the probability for offspring generation, we ensure that the evolution process is guided by enhancing the consistency of variable selection while eliminating non-informative variables. Finally, the choice of the final informative variable subset is based on the frequency of each variable being selected into the library of optimal chromosome. In this way, a variable of lower stability by itself yet still informative when combined with other variables would be included and evaluated.

Five case studies using different NIR datasets confirmed our hypothesis. These case studies show that CEEVS delivered the best variable selection consistency. They also show that CEEVS-based PLS soft sensor achieved the best prediction performance, when compared to GA, CARS or SVP based PLS soft sensor or the full PLS soft sensor without variable selection. More importantly, we show that CEEVS is able to identify the underlying chemical information, i.e., the wavelengths corresponding to the chemical bounds or functional groups that determine the sample properties of interest. In addition, CEEVS is not sensitive to its four tuning parameters when they are large enough, which is demonstrated by the fact that the same fixed parameters were used for all five case studies. The robustness of CEEVS to the tuning parameters corroborates with the findings that CEEVS has the highest variable selection consistency and the selected wavelengths correspond to important chemical bounds or functional groups. This robust performance is highly desirable, because it significantly simplifies tuning of the algorithm, and makes the implementation of CEEVS much easier than GA.

ASSOCIATED CONTENT

Supporting Information

S1: Additional figures showing the effectiveness of CEEVS in extracting chemical information from NIR spectra.

AUTHOR INFORMATION

Corresponding Author

Jin Wang – Department of Chemical Engineering, Auburn University, Auburn, AL 36849 USA, https://orcid.org/0000-0002-7638-8537; Email: wang@auburn.edu

Q. Peter He – Department of Chemical Engineering, Auburn University, Auburn, AL 36849 USA, https://orcid.org/0000-0002-2474-5950; Email: qhe@auburn.edu

Author

Jangwon Lee – Department of Chemical Engineering, Auburn University, Auburn, AL 36849 USA

Jesus Flores-Cerrillo – Linde Digital, Linde plc, Tonawanda, NY, 14150

Notes

The authors declare no competing financial interest.

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.

REFERENCE

(1) Tamburini, E.; Vaccari, G.; Tosi, S.; Trilli, A. Near-Infrared Spectroscopy: A Tool for Monitoring Submerged Fermentation Processes Using an Immersion Optical-Fiber Probe. *Appl. Spectrosc.* **2003**, *57*, 132–138.

(2) Shah, D.; Wang, J.; He, Q. P. A Feature-Based Soft Sensor for Spectroscopic Data Analysis. *J. Process Control* **2019**, *78*, 98–107.

(3) Li, H.; Liang, Y.; Xu, Q.; Cao, D. Key Wavelengths Screening Using Competitive Adaptive Reweighted Sampling Method for Multivariate Calibration. *Anal. Chim. Acta* **2009**. https://doi.org/10.1016/j.aca.2009.06.046.

(4) Hopkins, D. W. Shoot-out 2002: Transfer of Calibration for Content of Active in a Pharmaceutical Tablet. *NIR news* **2003**, *14* (5), 10–13.

(5) Wang, Z.; He, Q. P.; Wang, J. Comparison of Variable Selection Methods for PLS-Based Soft Sensor Modeling. *J. Process Control* **2015**, *26* (2015), 56–72. https://doi.org/10.1016/j.jprocont.2015.01.003.

(6) Andersen, C. M.; Bro, R. Variable Selection in Regression---a Tutorial. *J. Chemom.* **2010**, *24* (11–12), 728–737.

(7) Chong, I.-G. G.; Jun, C.-H. H. Performance of Some Variable Selection Methods When Multicollinearity Is Present. *Chemom. Intell. Lab. Syst.* **2005**, *78* (1–2), 103–112.

(8) Gosselin, R.; Rodrigue, D.; Duchesne, C. A Bootstrap-VIP Approach for Selecting Wavelength Intervals in Spectral Imaging Applications. *Chemom. Intell. Lab. Syst.* **2010**, *100* (1), 12–21.

(9) Kump, P.; Bai, E. W.; Chan, K. S.; Eichinger, B.; Li, K. Variable Selection via RIVAL (Removing Irrelevant Variables amidst Lasso Iterations) and Its Application to Nuclear Material Detection. *Automatica* **2012**, *48* (9), 2107–2115.

(10) Wold, S.; Johansson, E.; Cocchi, M. 3D-QSAR in Drug Design, Theory, Methods, and Applications. *ESCOM Science, Ledien*. 1993, pp 523–550.

(11) Centner, V.; Massart, D. L.; De Noord, O. E.; De Jong, S.; Vandeginste, B. M.; Sterna, C. Elimination of Uninformative Variables for Multivariate Calibration. *Anal. Chem.* **1996**, *68* (21), 3851–3858.

(12) Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. **1996**, *58*, 267–288.

(13) Leardi, R. Genetic Algorithms in Chemistry. *Journal of Chromatography A*. 2007, pp 226–233.

(14) Leardi, R.; Lupiáñez González, A. Genetic Algorithms Applied to Feature Selection in PLS Regression: How and When to Use Them. *Chemom. Intell. Lab. Syst.* **1998**, *41* (2), 195–207.

(15) Leardi, R. Application of Genetic Algorithm-PLS for Feature Selection in Spectral Data Sets. In *Journal of Chemometrics*; 2000; Vol. 14, pp 643–655.

(16) Chen, J.; Yang, C.; Zhu, H.; Li, Y.; Gui, W. A Novel Variable Selection Method Based on Stability and Variable Permutation for Multivariate Calibration. *Chemom. Intell. Lab. Syst.* **2018**, *182*, 188–201.

(17) Balabin, R. M.; Smirnov, S. V. Variable Selection in Near-Infrared Spectroscopy: Benchmarking of Feature Selection Methods on Biodiesel Data. *Anal. Chim. Acta* **2011**, *692* (1–2), 63–72.

(18) Xiaobo, Z.; Jiewen, Z.; Povey, M. J. W.; Holmes, M.; Hanpin, M.; Zou, X.; Zaho, J.; Povey, M. J. W.; Holmes, M.; Mao, H. Variables Selection Methods in Near-Infrared Spectroscopy. *Anal. Chim. Acta* **2010**, *667* (1–2), 14–32. https://doi.org/10.1016/J.ACA.2010.03.048.

(19) Tan, B.; Liang, Y.; Yi, L.; Li, H.; Zhou, Z.; Ji, X.; Deng, J. Identification of Free Fatty Acids Profiling of Type 2 Diabetes Mellitus and Exploring Possible Biomarkers by GC-MS Coupled with Chemometrics. *Metabolomics* **2010**, *6* (2), 219–228.

(20) Wu, D.; Sun, D. W. Application of Visible and near Infrared Hyperspectral Imaging for Non-Invasively Measuring Distribution of Water-Holding Capacity in Salmon Flesh. *Talanta* **2013**, *116*, 266–276.

(21) Fan, W.; Shan, Y.; Li, G.; Lv, H.; Li, H.; Liang, Y. Application of Competitive Adaptive Reweighted Sampling Method to Determine Effective Wavelengths for Prediction of Total Acid of Vinegar. *Food Anal. Methods* **2012**, *5* (3), 585–590.

(22) Xu, D.; Fan, W.; Lv, H.; Liang, Y.; Shan, Y.; Li, G.; Yang, Z.; Yu, L. Simultaneous Determination

of Traces Amounts of Cadmium, Zinc, and Cobalt Based on UV-Vis Spectrometry Combined with Wavelength Selection and Partial Least Squares Regression. *Spectrochim. Acta - Part A Mol. Biomol. Spectrosc.* **2014**, *123*, 430–435.

(23) Zheng, K.; Li, Q.; Wang, J.; Geng, J.; Cao, P.; Sui, T.; Wang, X.; Du, Y. Stability Competitive Adaptive Reweighted Sampling (SCARS) and Its Applications to Multivariate Calibration of NIR Spectra. *Chemom. Intell. Lab. Syst.* **2012**, *112*, 48–54.

(24) Corn dataset. http://www.eigenvector.com/data/Corn/index.html.

(25) Diesel Fuels dataset. http://eigenvector.com/data/SWRI/index.html.

(26) Pharmaceutical dataset. http://www.idrc-chambersburg.org/shootout_2002.html.

(27) Wheat dataset. https://www.wiley.com/legacy/wileychi/chemometrics/datasets.html.

(28) Nørgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B.; Norgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; et al. Interval Partial Least-Squares Regression (IPLS): A Comparative Chemometric Study with an Example from near-Infrared Spectroscopy. *Appl. Spectrosc.* **2000**, *54* (3), 413–419. https://doi.org/10.1366/0003702001949500.

(29) Bin, J.; Ai, F.; Fan, W.; Zhou, J.; Li, X.; Tang, W.; Liang, Y. An Efficient Variable Selection Method Based on Variable Permutation and Model Population Analysis for Multivariate Calibration of NIR Spectra. *Chemom. Intell. Lab. Syst.* **2016**, *158*, 1–13.

(30) Schulz, E.; Speekenbrink, M.; Krause, A. A Tutorial on Gaussian Process Regression: Modelling, Exploring, and Exploiting Functions. *J. Math. Psychol.* **2018**, *85*, 1–16.