ResearchGate

See discussions, stats, and author profiles for this publication at: https://www.researchgate.net/publication/321186811

A selective review and comparison for interval variable selection in spectroscopic modeling

Article *in* Chemometrics and Intelligent Laboratory Systems · January 2018

DOI: 10.1016/j.chemolab.2017.11.008

CITATIONS
0

READS
90

7 authors, including:

Li-Li Wang
Central South University
6 PUBLICATIONS **5** CITATIONS

SEE PROFILE

Nan Xiao
Central South University
39 PUBLICATIONS **76** CITATIONS

SEE PROFILE

Hong-Dong Li
Institute for Systems Biology
67 PUBLICATIONS **1,490** CITATIONS

SEE PROFILE

Qingsong Xu
Central South University
202 PUBLICATIONS **3,507** CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:

Project High dimensional survial data View project

Project Ordered Homogeneity Pursuit Lasso View project

All content following this page was uploaded by Nan Xiao on 15 February 2018.

The user has requested enhancement of the downloaded file.

Chemometrics and Intelligent Laboratory Systems 172 (2018) 229–240

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

A selective review and comparison for interval variable selection in spectroscopic modeling

Li-Li Wang [a,1], You-Wu Lin [b,1], Xu-Fei Wang [c], Nan Xiao [a,d], Yuan-Da Xu [e], Hong-Dong Li [f], Qing-Song Xu [a,*]

[a] School of Mathematics and Statistics, Central South University, Changsha 410083, PR China
[b] School of Mathematics and Statistics, Guangxi Teachers Education University, Nanning 530023, PR China
[c] Department of Statistics, Harvard University, 1 Oxford Street, Cambridge, MA 02138, USA
[d] Seven Bridges Genomics, 1 Main Street, Cambridge, MA 02142, USA
[e] Program of Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA
[f] School of Information Science and Engineering, Central South University, Changsha 410083, PR China

ARTICLE INFO

Keywords:
Spectroscopy
PLS
Interval variable selection

ABSTRACT

Dimension reduction and variable selection are two types of effective methods that deal with high-dimensional data. In particular, variable selection techniques are of wide-spread use and essentially consist of individual selection methods and interval selection methods. Given the fact that the vibrational spectra have continuous features of spectral bands, interval selection instead of individual spectral wavelength point selection allows for more stable models and easier interpretation. Numerous methods have been suggested for interval selection recently. Therefore, this paper is devoted to a selective review on interval selection methods with partial least squares (PLS) as the calibration model. We described the algorithms in the five classes: classic methods, penalty-based, sampling-based, correlation-based, and projection-based methods. Finally, we compared and discussed the performances of a subset of these methods on three real-world spectroscopic datasets.

1. Introduction

In recent years, the extensive use of multivariate calibration methods in multi-component spectral analysis has made them extremely popular techniques, especially for vibrational spectral data such as infrared (IR) spectroscopy, near infrared (NIR) spectroscopy, and ultraviolet–visible spectroscopy [1]. Multivariate calibration is devoted to the establishment of calibration models that relate variables to the properties of interest such as concentration values. Notably, wavelengths (spectral points) of the spectra are treated as variables in the modeling.

With the modern spectroscopic instrumental technology, a common feature of the obtained data is that there tend to be numerous variables but measured on much fewer samples, which is known as the challenging "large p, small n" problems in statistics. Take the NIR data investigated in this review for example. The spectra used for empirical analysis ranges within 10000–4000 cm^{-1} with an interval of 4 cm^{-1} yielding 1557 variables with only 67 samples involved. Such high-dimensional data raises the "curse of dimensionality" [2,3] that many traditional statistical methods cannot deal with [3,4]. To tackle the potential problems, two types of methods have been developed: dimension reduction and variable selection. The dimension reduction methods substitute the original high-dimensional variable space with relatively low-dimension spaces. The variable selection methods are dedicated to selecting important variables. Both types of the methods try to reduce the dimensionality of the original space and remove redundant variables while keeping the useful information of the original space as much as possible.

Partial least squares (PLS) [5–8] is a widely-used dimension reduction method based on latent variables and has gained extensive attention in a variety of fields such as chemometric, biomedicine, spectroscopy, and so forth [9,10]. It substitutes the feature space with the relatively low-dimension projected space, of which the direction is determined by latent variables consisting of combinations of the original variables. Despite the enhanced model accuracy, PLS is far from being perfect as its weak interpretability. Besides, although PLS reduces the model error caused by redundant and noisy variables, it is unable to cut them out directly and thoroughly.

* Corresponding author.
E-mail address: qsxu@csu.edu.cn (Q.-S. Xu).
[1] These authors contributed equally to this paper.

https://doi.org/10.1016/j.chemolab.2017.11.008
Received 29 May 2017; Received in revised form 7 September 2017; Accepted 2 November 2017
Available online 20 November 2017
0169-7439/© 2017 Elsevier B.V. All rights reserved.

Table 1
Results of different methods on the milk dataset. Statistical results with the form mean value ± standard deviation in 50 runs.

Methods	nLV	nVAR	RMSEP	RMSEC
PLS	9.8±0.6	1557.0±0.0	0.0448±0.0146	0.0142±0.0025
iPLS	6.6±2.2	145.6±58.3	0.0457±0.0251	0.0153±0.0135
MWPLS	8.1±1.7	279.1±183.5	0.0411±0.0092	0.0125±0.0053
EN-PLS	4.2±1.3	29.6±16.3	0.0659±0.0245	0.0292±0.0055
SIS-iPLS	7.1±1.7	212.3±62.0	0.0752±0.0253	0.0192±0.0082
FOSS	7.6±2.1	116.1±185.2	0.0436±0.0125	0.0105±0.0040

Researchers have verified both theoretically and experimentally that greater improvement of model performance can be achieved by variable selection [11–13]. In the context of spectroscopy, variable selection refers to identifying informative wavelengths (important regions to explain the information of the response variable) out of the full spectrum for the subsequent modeling. With the removal of the irrelevant and uninformative wavelengths, we can obtain a much simpler model without compromising its predictive ability [14]. By merits of the wavelength selection, a range of methods have been developed and can be grouped into two categories: individual wavelength selection methods and interval selection methods.

The representative individual wavelength selection methods include classic stepwise methods, e.g. forward selection [15], backward selection [16] and stepwise selection [17]; variable ranking-based strategy, e.g. loading weights [18], regression coefficients [19] and variable importance in projection (VIP) [20]; penalty-based strategy, e.g. least absolute shrinkage and selection operator (LASSO) [21], smoothly clipped absolute deviation (SCAD) [22,23] and sparse partial least squares (sPLS) [24]; model population analysis (MPA) [25] based strategy, e.g. random frog (RF) [26], iteratively retains informative variables (IRIV) [27], variable iterative space shrinkage approach (VISSA) [28] and bootstrapping soft shrinkage (BOSS) [29]; heuristic algorithm based strategy, e.g. simulated annealing (SA) [30], artificial neural networks (ANN) [31], genetic algorithm (GA) [32]; and some other methods, e.g. successive projection algorithm (SPA) [33], uninformative variable elimination (UVE) [34] and UVE-SPA method [35]. For spectroscopic data, since functional groups absorb within relatively short wavelength bands, continuous and adjacent wavelengths are highly correlated. Wavelengths with high correlation tend to contain shared information and have identical regression coefficients [36]. Therefore, models established on any one of the correlated variables are expected to perform similarly [37]. Thus, in turn, can make it difficult to determine the important variables and impede the interpretability of models. Studies have shown that performances of calibration models based on wavelength intervals tend to be more robust than that based on individual wavelengths [13]. Besides, the vibrational spectral band relating to chemical band generally has a width of 4–200 cm^{-1} . So, the selection of spectral intervals not only can provide reasonable interpretation, but also makes more sense and is expected for the best performance. Inspired by the advantages of interval selection, numerous methods have been proposed and developed.

This review highlights the interval selection methods for spectroscopic data and is organized as follows. Because we take PLS as the calibration model method, a commonly used algorithm for PLS is first presented in Section 2. Section 3 reviews the theories and algorithms for a selective set of interval selection methods, which are organized into five categories: classic methods including interval PLS (iPLS) [38] and its variants, moving windows PLS (MWPLS) [39] and its variants; penalty-based methods including elastic net combined with partial least squares regression (EN-PLSR) [40], iterative rank PLS regression coefficient screening (EN-IRRCS) [41], and group PLS (gPLS) [42]; sampling-based methods including iPLS-Bootstrap [43], Bootstrap-VIP [44], Fisher optimal subspace shrinkage (FOSS) [45], interval random frog (iRF) [46], and interval variable iterative space shrinkage approach (iVISSA) [37]; correlation-based method including SIS-iPLS [47] and projection-based method including interval successive projections algorithm (iSPA) [48]. Section 4 describes three near-infrared spectroscopy datasets and software used for the evaluation of six methods, PLS, iPLS, MWPLS, EN-PLS, SIS-iPLS, and FOSS. Experimental results are shown and discussed in Section 5, followed by the summary in Section 6.

2. Partial least squares

Since PLS is used for building the calibration model in this work, a brief description of PLS is provided in this section.

PLS constructs linear relations between response variables and predictors using latent variables comprised of combinations of the predictors. It breaks the high-dimensional data down into scores and loadings determined by both response variables and predictors. A variety of PLS algorithms are available, such as PLS1 [49], PLS2 [50], PLS-SB [51], SIMPLS [52], and GPLS [53]. For simplicity, this section focuses on the linear model with one single response variable known as the PLS1 algorithm.

Consider the multiple linear regression model.

$$\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\varepsilon} = \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \ldots + \boldsymbol{x}_p\beta_p + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{X}$ is an n × p matrix containing p features of the collected data, $\mathbf{y}$ of size n × 1 is the property of interest for a set of n samples, $\boldsymbol{\beta}$ is a vector of unknown parameters, $\boldsymbol{\varepsilon}$ is an error term with mean zero and variance $\sigma^2\boldsymbol{I}$. Variables are centered to have mean zero before any further operations. In the PLS1 algorithm, $\mathbf{X}$ is decomposed into the score vectors (latent variables):

$$\mathbf{X} = \mathbf{T}U' + E \tag{2}$$

where $\mathbf{T}$ is an n × A matrix of A latent variables (also called scores), the p × A matrix $\mathbf{U}$ represents A loading vectors for $\mathbf{X}$, and the n × p matrix $\mathbf{E}$ is the residual. Particularly, latent vectors in $\mathbf{T}$ are linear combinations of the original predictors:

$$\mathbf{T} = (t_1, \ldots, t_A) = \mathbf{XW} = \left(\sum_{j=1}^{p} \boldsymbol{x}_j w_{j1}, \ldots, \sum_{j=1}^{p} \boldsymbol{x}_j w_{jA}\right) \tag{3}$$

where the p × A matrix $\mathbf{W} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_A)$ represents the weight vectors for A latent variables. Then the response is projected to the latent variables space:

$$\mathbf{y} = \mathbf{Tq} + \mathbf{f} = T(T'T)^{-1}T'y + \mathbf{f} \tag{4}$$

where the A × 1 vector $\mathbf{q}$ is loading vector for $\mathbf{y}$, the n × 1 vector $\mathbf{f}$ represents the residual. Particularly, $\mathbf{q}$ is the least squares estimation by regressing $\mathbf{y}$ against the score matrix $\mathbf{T}$. Thus,

$$\widehat{\mathbf{y}} = \mathbf{Tq} = \mathbf{XWq} = XW(T'T)^{-1}T'y = \mathbf{X}\widehat{\beta}_{PLS}. \tag{5}$$

So, the partial least squares estimator can be written as:

$$\widehat{\beta}_{PLS} = W(T'T)^{-1}T'y = W(W'X'XW)^{-1}W'X'y \tag{6}$$

3. Interval selection methods

This section describes the theories and algorithms of some interval selection methods. To make it easier to read and understand, we classify these methods into five categories: classic methods, penalty-based methods, sampling-based methods, correlation-based methods, and projection-based methods.

3.1. Classic interval selection methods

3.1.1. Interval PLS (iPLS) and its variants

A representative approach of interval selection is interval partial least

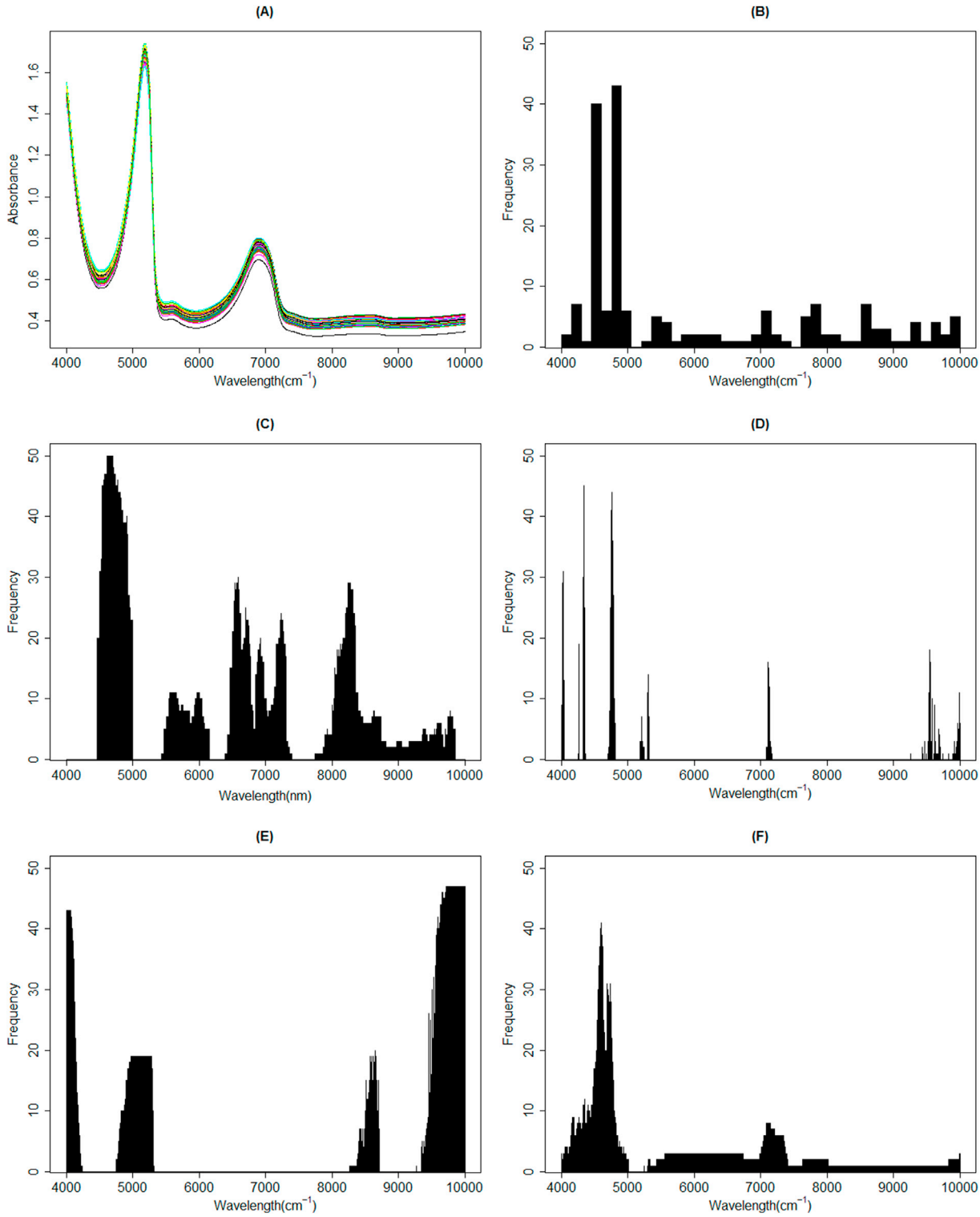

Fig. 1. Frequency of the selected variables by different methods in 50 runs on the milk dataset. (A) PLS; (B) iPLS; (C) MW-PLS; (D) EN-PLS; (E) SIS-iPLS; (F) FOSS.

squares (iPLS) [38] proposed by Norgaard et al., which splits the full spectrum into equal-width and non-overlapping sub-intervals and establish local PLS models of the same dimensionality in each interval. To find the best interval explaining the information of the response variable, the predictive performances of both the local models and the model built on the whole spectrum are compared based on the root mean squared error of cross-validation (RMSECV). There is not much possibility for this sub-interval to hit the optimal interval because the widths of the split intervals are equal and they are not overlapping with each other. Therefore, some simple optimization can be carried out to refine the interval limits. The optimization mainly comprises two steps, including (1) interval shift; (2) changes in interval width: two-sided (symmetrical), one-sided (asymmetrical, left), or one-sided (asymmetrical, right).

iPLS gives a first impression of the information of different sub-intervals and locates the individual best interval. However, it fails to take the synergism among intervals into consideration and thus obtain a

Table 2
Results of different methods on the tobacco dataset. Statistical results with the form mean value ± standard deviation in 50 runs.

Methods	nLV	nVAR	RMSEP	RMSEC
PLS	12.0±0.0	1557.0±0.0	0.0067±0.0005	0.0039±0.0001
iPLS	12.0±0.0	208.3±30.1	0.0049±0.0003	0.0025±0.0001
MWPLS	11.8±0.5	179.4±37.6	0.0058±0.0005	0.0032±0.0002
EN-PLS	9.5±1.4	29.1±5.6	0.0081±0.0006	0.0048±0.0002
SIS-iPLS	10.4±2.0	139.2±72.1	0.0072±0.0025	0.0040±0.0016
FOSS	12.0±0.0	177.8±40.6	0.0049±0.0004	0.0026±0.0001

suboptimal model. So, the combination of multiple intervals may lead to a PLS model with better performance. Synergy interval partial least squares (siPLS) [54] and backward (forward) interval partial least squares (biPLS/fiPLS) [55,56] are three primary extensions of iPLS that consider different combinations of intervals based on iPLS. SiPLS is a strategy that searches for a proper combination of intervals among all the combinations of two, three, and four intervals. And the combination with the lowest RMSECV is regarded as the optimized intervals. In comparison, biPLS and fiPLS couple iPLS with backward and forward selection procedures, respectively. The general algorithms work in two steps, including dividing the spectrum into multiple intervals the same way as iPLS and removing (adding) one interval at a time whose removal (addition) results in the lowest RMSECV.

3.1.2. Moving window PLS (MWPLS) and its variants

One of the main features of iPLS is that adjacent and non-overlapping sub-divisions of the whole spectrum are tested using PLS models with identical dimensionality. This fact may increase the possibility of missing important intervals and enlarge the influence of the choice of dimensionality. Moving window partial least squares (MWPLS) [39] alleviates the aforementioned problems by moving along the spectrum with a fix-sized window and allowing varying dimensionalities for PLS models in each window. And the informative spectral regions are identified on the basis of low model complexity and a desired prediction error level. A final PLS model is built by including all informative intervals or built as an ensemble of local PLS models on each interval.

It is worth noting that sub-regions of the informative regions selected by MWPLS may provide better models than the original fixed-size regions. Thus, a further sampling search of the optimal sub-region within the informative interval seems to be necessary in order to reach a better model performance. Changeable size moving window partial least squares (CSMWPLS) [57] is a strategy that collects all possible sub-intervals within an interval by changeable size windows and then determines the optimal sub-region as the one with the lowest RMSECV. Searching combination moving window partial least squares (SCMWPLS) [57] exploits a new strategy to optimize the combination of intervals based on CSMWPLS using the forward selection procedure.

3.2. Penalty-based interval selection methods

This group method contains a type of methods that select or leave out correlated variables together. The examples include the elastic net [58], group lasso [59], etc. The two methods succeed to choose groups of variables by implementing the penalty function. Therefore, it is reasonable to apply the group selection methods in the field of spectroscopy to pick informative intervals. In this section, we display three penalty-based interval selection methods involving EN-PLSR, EN-IRRCS, and group PLS.

3.2.1. Elastic net combined with partial least squares regression (EN-PLSR)

Lasso [21] is a variable selection method proposed by Tibshirani et al. which penalizes the loss function with L_1 -norm of coefficients. To make it clear, Lasso is stated as:

$$\widehat{\beta}^{Lasso} = arg\min_{\beta} \|\boldsymbol{y} - X\beta\|_2^2 + \lambda_1\|\beta\|_1 \tag{7}$$

where λ_1 is tuning parameter. L_1 -norm in the latter term of Equation (6) turns regression coefficients to exact zeros, which leads to the elimination of variables. However, Lasso treats variables as independent ones, which contradicts the continuity of spectroscopic data that consecutive wavelengths are highly correlated. Zou improved Lasso by adding a L_2 -penalty of coefficients and designated it as elastic net (EN) [58]:

$$\widehat{\beta}^{EN} = (1+\lambda_2)\left\{arg\min_{\beta} \|\boldsymbol{y} - X\beta\|_2^2 + \lambda_1\|\beta\|_1 + \lambda_2\|\beta\|_2^2\right\} \tag{8}$$

where λ_1 and λ_2 are tuning parameters. EN not only remains the ability to filter unimportant variables out, but also enables to estimate the regression coefficients of strongly correlated variables with close absolute values. So, linearly correlated variables within an interval are selected in or left out together, which is known as the grouping effect.

Fu et al. proposed an interval selection technique called elastic net combined with partial least squares regression (EN-PLSR) [40]. It generally contains two phases, including identifying important intervals by elastic net and further screening informative intervals by the recursive leave-one-group-out strategy. Its ability to select important intervals comes from two aspects. First, EN provides a way to filter out unnecessary variables. The second aspect follows the unique feature of spectroscopic data that strong correlations exist among successive wavelengths. Thus, the grouping effect of EN makes it possible to select important consecutive wavelengths together. The specific description of EN-PLSR is demonstrated in four steps.

(1) Apply the elastic net to the whole spectra and suppose the remaining variable sequence constructs m intervals.
(2) Build PLS models on $(m-1)$ intervals with one interval left out sequentially and compute the values of RMSECV.
(3) Delete the interval associated with the lowest RMSECV value.
(4) Repeat step (2)–(3) until the lowest RMSECV in every iteration starts to increase. The remaining intervals are considered the selected informative variables.

3.2.2. Elastic net based on iterative rank PLS regression coefficient screening (EN-IRRCS)

Motivated by the grouping effect of the elastic net, Huang developed a method for interval selection designated as elastic net based on iterative rank PLS regression coefficient screening (EN-IRRCS) [41] which couples the elastic net with the technique of ranking PLS coefficients. EN-IRRCS first employs the rank of PLS regression coefficients to eliminate a portion of variables. On this basis, EN is next used to filter more variables out. And the two-step screening procedure is carried out iteratively to include more variables and to mitigate the risk of missing important variables. We give more details about EN-IRRCS in five steps.

(1) Build a PLS model on the variable space and sort the absolute regression coefficients in decreasing order. Select and record variables with the top k largest absolute coefficients.
(2) Apply elastic net on the k variables to further extract a subset of intervals, denoted by M.
(3) Update the response with the residual vector of regressing y against M. Take all variables but the selected interval subset M as the new variable space.
(4) Repeat step (1)–(3) till the size of the union of the disjoint interval subsets obtained in every iteration is less than that of samples. And consider the union as the optimal intervals.

It is clear that larger coefficients indicate stronger relations with the response. Thus, it is reasonable to regard the wavelengths with small PLS regression coefficients as uninformative ones and discard them. Therefore, EN-IRRCS is able to filter out a proportion of the unimportant variables using regression coefficients. Moreover, for spectroscopic data, successive wavelengths tend to behave highly correlated and have close

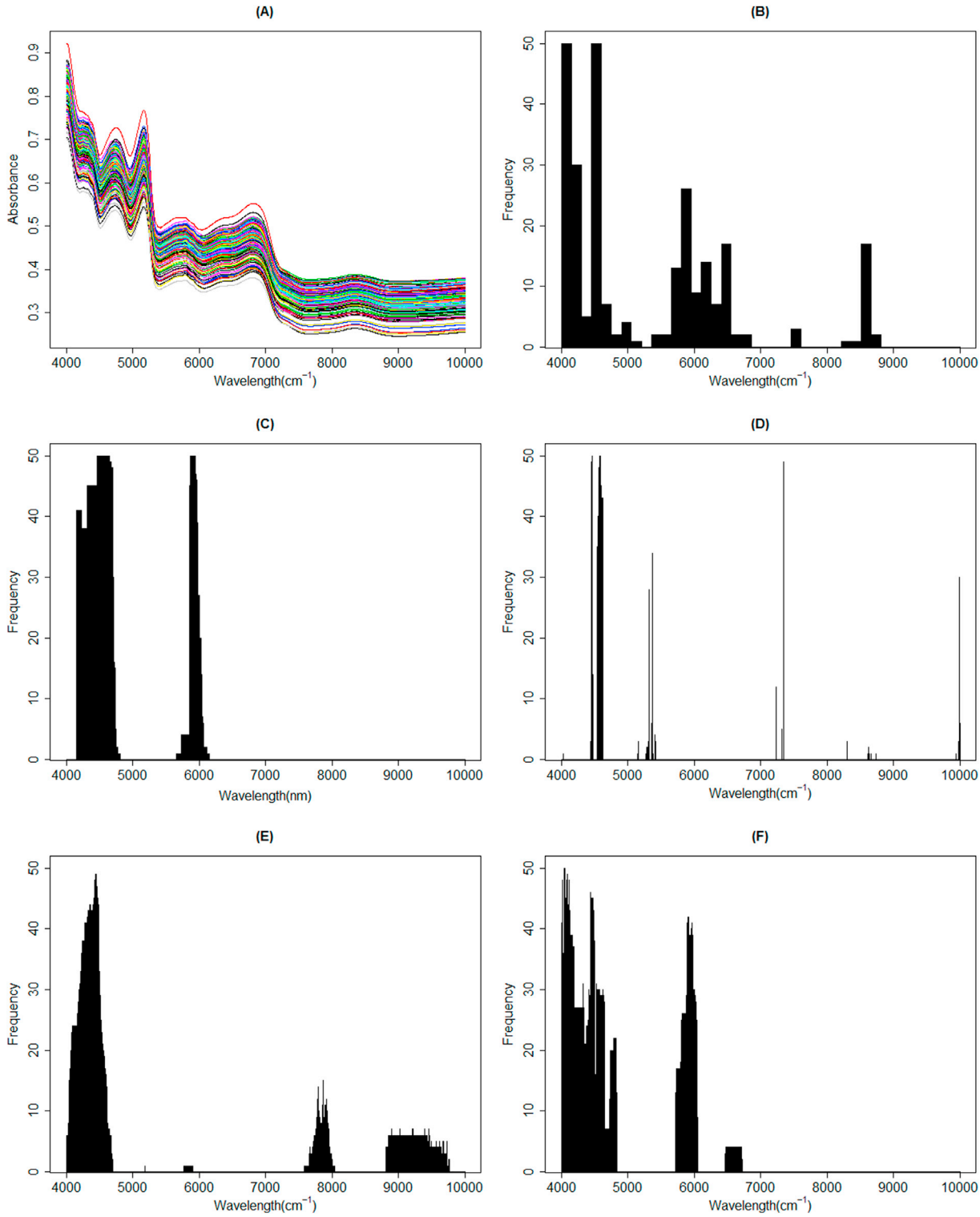

Fig. 2. Frequency of the selected variables by different methods in 50 runs on the tobacco dataset. (A) PLS; (B) iPLS; (C) MW-PLS; (D) EN-PLS; (E) SIS-iPLS; (F) FOSS.

regression coefficients in PLS model. So this property guarantees EN-IRRCS to select consecutive wavelengths. EN-IRRCS filters more unnecessary variables out by applying the elastic net. Notably, the elastic net picks out intervals again owing to its grouping effect. However, the two-phase variable space shrinkage may exclude some informative wavelengths as well. To include relevant wavelengths possibly missed in the two-phase shrinkage, screening procedures are carried out iteratively to gain unions of disjoint important interval sets derived in each iteration.

3.2.3. Group PLS (gPLS)

EN-PLS and EN-IRRCS both generate intervals by the grouping effect of EN derived from L_2 -penalty. Group PLS (gPLS) [42] by comparison, realizes interval selection under the framework of optimization, where the group lasso penalty [59] is imposed on the Frobenius-norm loss

Table 3
Results of different methods on the soil SOM dataset. Statistical results with the form mean value ± standard deviation in 50 runs.

Methods	nLV	nVAR	RMSEP	RMSEC
PLS	10.0±0.0	700.0±0.0	0.5050±0.0777	0.2499±0.0130
iPLS	9.1±1.9	137.3±49.4	0.5203±0.2008	0.2615±0.1342
MWPLS	9.9±0.3	158.1±32.7	0.3442±0.0524	0.1742±0.0203
EN-PLS	8.7±1.7	54.8±32.2	0.7995±0.1260	0.4214±0.0471
SIS-iPLS	9.3±0.8	119.7±3.8	0.4600±0.0886	0.2363±0.0224
FOSS	9.7±0.8	110.8±108.7	0.3371±0.0758	0.1436±0.0191

Table 4
Abbreviations used in this paper.

PLS	Partial Least Squares
LASSO	Least Absolute Shrinkage and Selection Operator
sPLS	Sparse Partial Least Squares
iPLS	Interval Partial Least Squares
siPLS	Synergy Interval Partial Least Squares
biPLS/ fiPLS	Backward/Forward Interval Partial Least Squares
MWPLS	Moving Windows Partial Least Squares
CSMWPLS	Changeable Size Moving Window Partial Least Squares
SCMWPLS	Searching Combination Moving Window Partial Least Squares
EN	Elastic Net
EN-PLSR	Elastic Net combined with Partial Least Squares Regression
EN-IRRCS	Elastic Net based on Iterative Rank PLS Regression Coefficient Screening
gPLS	Group Partial Least Squares
BOSS	Bootstrapping Soft Shrinkage
FOSS	Fisher Optimal Subspace Shrinkage
WBBS	Weighted Block Bootstrap Sampling
FOP	Fisher Optimal Partition
RF	Random Frog
iRF	Interval Random Frog
MPA	Model Population Analysis
VISSA	Variable Iterative Space Shrinkage Approach
iVISSA	Interval Variable Iterative Space Shrinkage Approach
VIP	Variable Important in Projection
SIS	Sure Independence Screening
SIS-iPLS	Sure Independence Screening and Interval Partial Least Squares
SPA	Successive Projections Algorithm
iSPA	Interval Successive Projections Algorithm
RMSEP	Root Mean Squares Error of Prediction
RMSEP	Root Mean Squares Error of Calibration

function. The optimization problem can be written as:

$$\min_{u_a,v_a}\left\{\sum_{g=1}^{G}\sum_{l=1}^{L}\left\|\boldsymbol{C}_a^{(g,l)}-u_a^{(g)}v_a^{(l)^T}\right\|_F^2+\lambda_1\sum_{g=1}^{G}\sqrt{p_g}\left\|u_a^{(g)}\right\|_2+\lambda_2\sum_{l=1}^{L}\sqrt{q_l}\left\|v_a^{(l)}\right\|_2\right\} \tag{9}$$

where variables (wavelengths) of $\mathbf{X}$ ($\mathbf{y}$) are divided into G (L) groups (intervals) by assembling columns of $\mathbf{X}$ ($\mathbf{y}$), $\boldsymbol{C}_a^{(g,l)}=X_a^{(g)^T}\boldsymbol{y}_a^{(l)}$, $u_a^{(g)}$ ($v_a^{(l)}$) is the PLS loading vector related to variables in the group g (l), p_g (q_l) indicates the variable number of group g(l), λ_1 and λ_2 are tuning parameters and a represents PLS dimension. In this paper, we focus on the case where $\dim(\mathbf{y})=1$ and thus $\mathrm{L}=1,\ q_l=1$. So, the optimization problem can be transformed to

$$\min_{u_a,v_a}\left\{\sum_{g=1}^{G}\left\|\left\lfloor\boldsymbol{C}_a^{(g)}-u_a^{(g)}v_a\right\rfloor\right\|_F^2+\lambda_1\sum_{g=1}^{G}\sqrt{p_g}\left\|u_a^{(g)}\right\|_2+\lambda_2\|v_a\|_2\right\} \tag{10}$$

The optimization function consists of the loss function and the penalty. The penalization term on the loading vector $u_a^{(g)}$ should be emphasized here because it makes gPLS an interval selection method. The penalty can be seen as the fusion of L_1 and L_2 penalty. L_1 penalty encourages sparsity on the G groups (intervals). On the other hand, the L_2 penalty, imposed on the loading vectors associated with different group variables, encourages the "grouping effect" within groups (intervals), which enables the method to put together the consecutive and related wavelengths. Therefore, the overall design of the gPLS penalty can yield sparsity on the group level, which indicates the selection of spectrum intervals.

Sparse PLS (sPLS) [60, 61], gPLS and sparse group PLS (sgPLS) [42] share the nature that they work on the predictor matrix decomposition taking into account sparsity in the data structure. So, a comparison of the three sparsity methods deserves to be mentioned. One remarkable difference lies in the optimization penalty and the induced sparseness. More specifically, sparse PLS uses L_1 -penalty to achieve sparsity in individual variables, while sgPLS is able to drop groups of variables and individual variables within groups simultaneously through the combination of L_1 -penalty and the penalty used in gPLS.

3.3. Sampling-based interval selection methods

The sampling techniques also widely appear in the process of developing interval methods. Commonly used sampling strategies primarily fall into two classes: sampling in the sample space and sampling in the variable space. For the two sampling types, a limited number of interval selection methods have been developed. The representative methods using bootstrap in the sample space include iPLS-Bootstrap and Bootstrap-VIP. Besides, FOSS, iRF, and iVISSA employ Monte Carlo, weighted binary matrix, and weighted block bootstrap as the sampling strategy in the variable space respectively, to select intervals. Other sampling methods [62] can also be incorporated for sampling-based interval selection methods.

3.3.1. iPLS-bootstrap

Bootstrap is a sampling technique proposed by Efron and often used for statistical inference [63]. It draws sub-datasets with replacement for multiple times. For each bootstrap sample, a sub-model is built and the statistic value of interest is estimated. Then the statistical analysis is based on the values of interest derived from all bootstrap samples. Examples using bootstrap involve the calculations of the mean, distribution, and confidence interval of concerned statistics.

PLS-Bootstrap [64], a variable selection method, takes advantage of the bootstrap applied in the confidence interval analysis, to eliminate unimportant variables. In PLS-Bootstrap, PLS models are first built on different bootstrap samples. Based on the different models, confidence intervals can be constructed for the PLS regression coefficients. The variable is considered uninformative if its confidence interval includes the value 0.

Considering the continuity of the spectroscopic data, scattered variables derived from PLS-Bootstrap may be suboptimal. So, interval PLS-Bootstrap (iPLS-Bootstrap) [43] is developed. It gives a way to transform the discrete variables obtained from PLS-Bootstrap into continuous and informative bands. More details about the algorithm are presented as follows.

(1) Apply PLS-Bootstrap on the wavelengths and obtain an important variable sequence.
(2) Assemble intervals based on the variables derived from PLS-Bootstrap. Define a wavelength to be the terminator of an interval, if the region between this wavelength and the next wavelength in the sequence contains more variables than a predetermined number.
(3) For assembled intervals, cut out those with fewer variables than a predefined threshold and those with a weak contribution to the PLS model (See details of the criterion in Ref. [43]).

iPLS-Bootstrap consists of two main steps including the construction of the preliminary bands and the removal of unnecessary intervals. The criterion for extending scattered informative variables to continuous

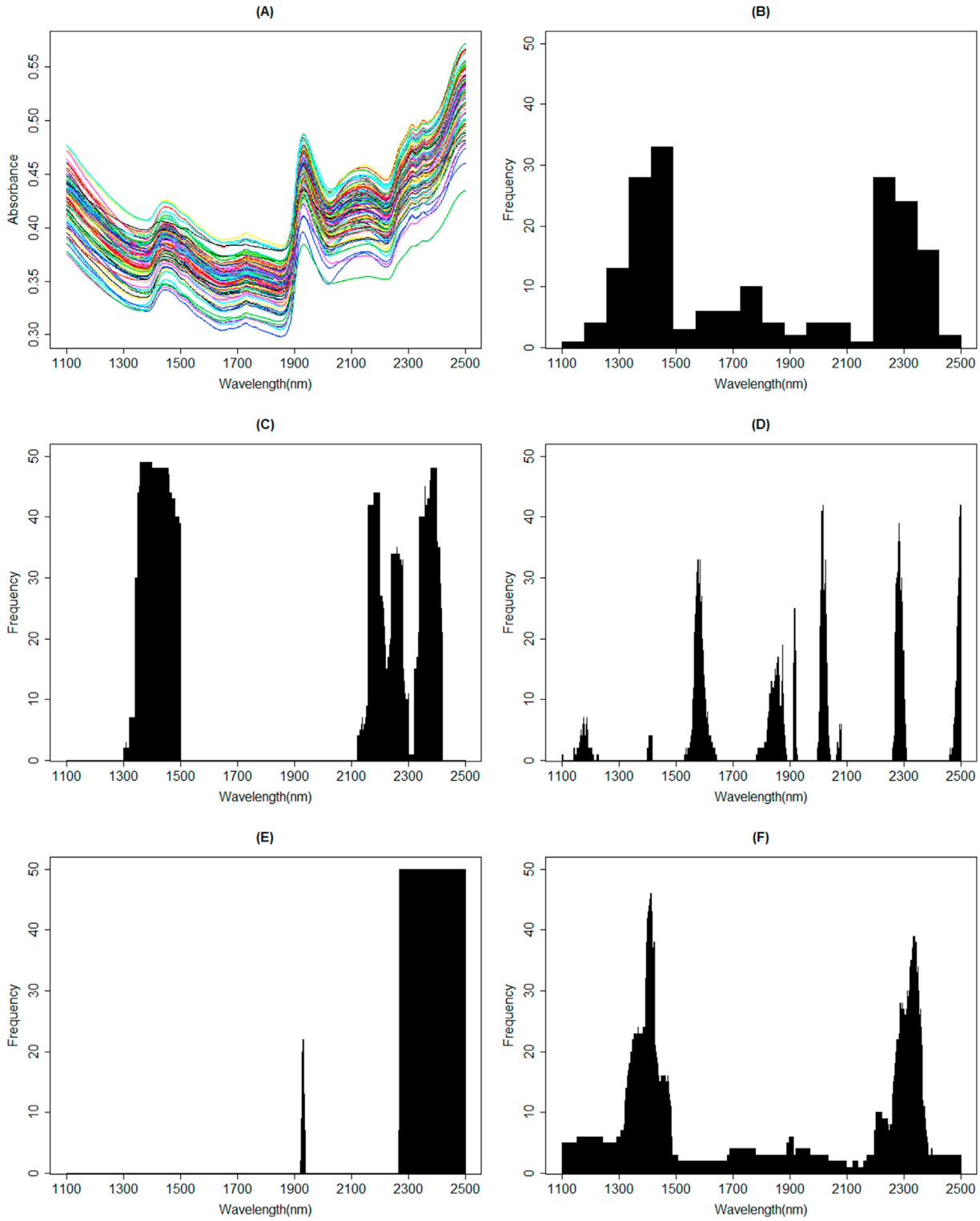

Fig. 3. Frequency of the selected variables by different methods in 50 runs on the soil SOM dataset. (A) PLS; (B) iPLS; (C) MW-PLS; (D) EN-PLS; (E) SIS-iPLS; (F) FOSS.

intervals in step (2) should be focused. It aims to make the two variables as the edge of an interval close enough so that the region between them keeps informative as well. The following screening ensures the selected intervals to contain sufficient information and great contribution to the response.

3.3.2. Bootstrap-VIP

PLS regression coefficients reflect the variable importance to some extent. Therefore, it makes sense that PLS-Bootstrap measures variable significance by applying bootstrap to construct confidence interval of the PLS regression coefficients. The variable importance on the projection (VIP) [20] is also a metric for assessing the variable significance. Thus, it is natural to make use of VIP for selecting variables.

Bootstrap-VIP [44] is a modified and robust version of VIP. It combines the bootstrap technique with VIP. Notably, bootstrap-VIP "is proposed as a simple wavelength selection method, yet having the ability to

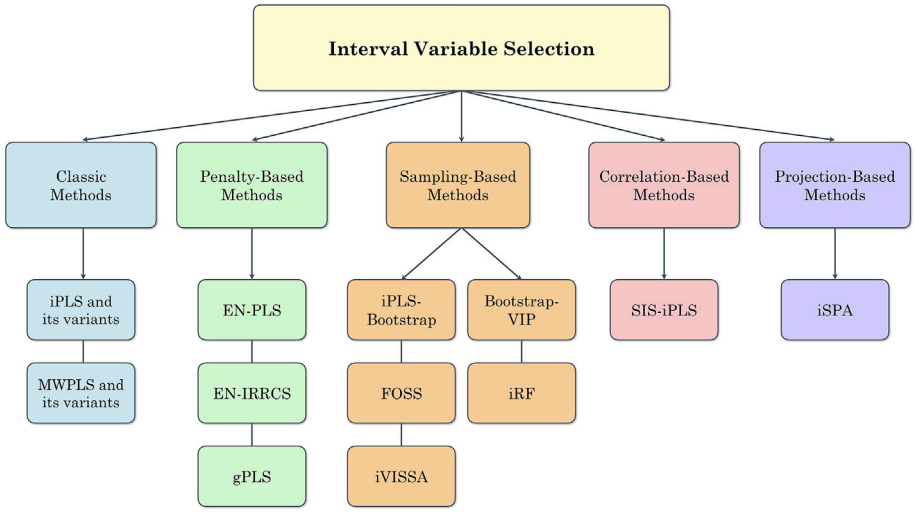

Fig. 4. An outline for interval selection methods reviewed in this paper.

identify relevant spectral intervals." [44] It works similarly as PLS-Bootstrap, except for two aspects. First, VIP rather than the regression coefficient is taken as the measure of variable importance and the statistic of interest for statistical analysis. The second distinction lies in the criterion for deciding informative variables. Specifically, a variable is considered relevant and informative if its average VIP value of all bootstrap samples is above 1.0. Compared to the VIP, Bootstrap-VIP improves the predictive accuracy and obtains more grouped intervals [44].

3.3.3. Fisher optimal subspace shrinkage (FOSS)

Interval PLS-Bootstrap and Bootstrap-VIP described above utilize bootstrap as the sampling strategy in the sample space. Empirical studies show that compared to the full-spectrum PLS and PLS-Bootstrap, they exhibit better predictive performance [43,44]. Recently, Deng et al. extended the bootstrap technique to the variable space and developed an effective variable selection method designated bootstrapping soft shrinkage (BOSS) [29]. However, bootstrap only suits the case of the independent data but not the highly collinear variables in spectroscopy.

To enhance the model performance of BOSS, Lin et al. developed an approach for wavelength interval selection designated Fisher optimal subspace shrinkage (FOSS) [45], which employs the weighted block bootstrap sampling (WBBS) [65] and Fisher optimal partition (FOP) [66] as the sampling and partition methods, respectively. The FOSS algorithm is described below.

(1) Build a PLS model over the variable space and compute the value of RMSECV.
(2) FOP is applied on the regression coefficients to construct sub-intervals.
(3) Calculate the mean of the absolute regression coefficients of the variables within every interval. And take it as the sampling weight for the interval.
(4) Resample the sub-intervals with replacement and use the sampling weights as the probability for intervals to be chosen.
(5) Update the variable space by the combination of drawn intervals.
(6) Repeat step (1)–(5) till only one variable is left to be sampled.
(7) Select the interval combination with the smallest RMSECV among all iterations as the optimal intervals for the best prediction performance.

What is worth pointing out is that FOP is applied based on the information of regression coefficients instead of observations as in conventional cases. Since consecutive wavelengths take close values of PLS coefficients [36], FOP is capable of dividing the spectra into continuous intervals and provide the foundation of the interval selection in following procedures. The WBBS algorithm applied in the variable space is particularly suitable for sampling consecutive wavelengths with high correlation. It provides a proper way to draw varied combinations of blocks of variables. Besides, WBBS enables the algorithm to shrink the variable space softly, because variable blocks with larger weights tend to be chosen with higher probability. Therefore, the seamless integration of FOP and WBBS makes FOSS a promising interval selection method for data with highly correlated variables, such as the data from spectroscopy, quantitative genetics, and metabolomics.

3.3.4. Interval random frog (iRF)

Monte Carlo [67] is also an extensively used technique for sampling. Similar to Bootstrap, it resamples datasets randomly without replacement for multiple times to generate a large number of sub-datasets. The random frog (RF) [26] algorithm conducts analysis on a large number of variable subsets using a reversible jump Markov Chain Monte Carlo (RJMCMC)-like strategy for single variable selection. Interval random frog (iRF) [46] modifies the original RF algorithm by using spectra intervals instead of wavelength points as the variables, which makes iRF an interval selection method. It is worth noting that the random sampling technique is applied to the variable space to assemble different combinations of intervals for building models. The explicit algorithm is described as follows.

(1) A moving window with a fixed width splits the whole spectra into overlapping sub-intervals. A set of m intervals, denoted by M_0, is sampled randomly from the interval pool.
(2) Generate a number m^* from the normal distribution $N(\mathrm{m}, \theta\mathrm{m})$ to specify the number of intervals to be selected into the candidate set M^*, where θ controls the quantitative range of the candidate set dimension.
(3) The candidate set is determined based on M_0 in two ways: (a) remove a certain number of intervals from M_0; (b) combine M_0 with some randomly drawn intervals from the interval pool. For more details, see Ref. [46].
(4) The candidate set M^* is accepted as the new interval set M_1 with a certain probability regarding their associated RMSECV values.
(5) Repeat step (2)–(4) N times and record M_i, $i = 1, 2, \ldots, \mathrm{N}$.
(6) Each interval is ranked according to the interval importance in the sense of its frequency in N iterations.

The normal distribution in step (2) controls the addition and deletion of intervals in the candidate set and allows for "great jumps between differently dimensioned models" and "the refinement of model dimensionality" [46]. The resampling strategy serves as a tool to extract different interval combinations. Therefore, a chain of interval sets is defined and subsequently evaluated based on RMSECV.

3.3.5. Interval variable iterative space shrinkage approach (iVISSA)

As described above, iRF adopts the Monte Carlo technique as the sampling strategy to draw sub-interval sets. Apart from Monte Carlo, the weighted binary matrix sampling (WBMS) [28] is raised as a sampling strategy and has been applied in the interval selection. Deng's work presented a wavelength selection method named variable iterative space shrinkage approach (VISSA) [28], where WBMS is employed as the sampling method in the variable space to extract a population of combinations of variables for the model population analysis (MPA) [25].

Interval variable iterative space shrinkage approach (iVISSA) [37] is a modified version of VISSA for selecting spectral intervals. Similar to VISSA, iVISSA implements the global search procedure by applying WBMS on the variables to identify the optimal locations and combinations of informative spectral wavelengths. The global search procedure serves the purpose of finding important individual wavelength points. The difference between iVISSA and VISSA lies in the additional local search procedure. It follows to extend the individual points to continuous intervals. Given the continuous nature of spectroscopic data, it is reasonable for the local search procedure to seek important variables near the informative individual wavelengths. Therefore, both of the procedures enable iVISSA to identify consecutive and informative intervals. The detailed iVISSA algorithm is:

(1) Construct a binary matrix of size $N \times p$ representing N subdatasets and p variables. The rows reflect the results of sampling, where 1 in a column indicates that the associated variable is used for modeling, and vice versa. The sampling weight $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_p)$ controls the frequency of 1 in each column.
(2) Build N PLS models on N data sets. Update the sampling weight ω_i using the frequency of the i -th variable in the first 10% models with the lowest RMSECV.
(3) If the weight turns into 1, the corresponding variable will be selected for modeling in all datasets in every iteration. Therefore, it is considered important and will be used for the final calibration model.
(4) Combine the important variable in step (3) with its adjacent variable one spectral point at a time. A series of PLS models are then built and assessed to search for the optimal interval width.
(5) Run step (1)–(4) iteratively until the sampling weight $\boldsymbol{\omega}$ stay constant.

3.4. Correlation-based interval selection method

3.4.1. SIS-iPLS

An interval selection method based on the correlation between the response variable and predictors was proposed by Xu et al. named sure independence screening and interval PLS (SIS-iPLS) [47]. It utilizes the correlation-based SIS [4] algorithm to sort wavelengths, then constructs preliminary intervals. The variable space is further shrunk by the backward selection for a better predictive performance. Before modeling, X and y are centered and scaled. We introduce the SIS-iPLS algorithm in the following five steps.

(1) Calculate the correlations of p predictors with the response. Sort the p correlations in the decreasing order and extract the first k variables with the largest correlations.
(2) Suppose the variable sequence obtained in step (1) constitutes m intervals. Note that consecutive variables are regarded as an interval.
(3) Establish m PLS models on the $(m-1)$ intervals with one interval removed in turn.
(4) The uninformative interval is defined as the one whose removal results in the lowest RMSEP.
(5) Repeat step (3)–(4) and eliminate one interval at a time.

As is known, one of the typical features of the spectroscopic data is the strong correlation among successive wavelengths. The adjacent variables tend to have close correlations with the response. Therefore, the threshold k enables to keep consecutive variables together, which allows for the preliminary construction of intervals. Additionally, SIS has been proved to enjoy the sure independence screening property, which tells that selected variables by SIS contain the true model with probability tending to 1. Therefore, it guarantees that variables surviving SIS tend to contain all informative variables. However, it is likely that except for the important variables, some uninformative variables are also included in the constructed intervals. To solve this problem, SIS-iPLS further employs a modified version of the stepwise backward variable selection algorithm. This algorithm eliminates one interval instead of a variable at a time until the optimal model performance is achieved. The preliminary intervals set retained from SIS and the following interval-wise elimination procedure allow SIS-iPLS to choose important intervals.

3.5. Projection-based interval selection method

3.5.1. Interval successive projections algorithm (iSPA)

A forward variable selection technique termed successive projections algorithm (SPA) [33] is proposed. SPA uses projection operations iteratively to alleviate the collinearity among variables and achieves good prediction ability. Based on the idea of SPA, the projection operation is further developed for interval selection. Interval successive projections algorithm (iSPA) [48] is proposed using projections to select informative intervals. The process of iSPA is as follows.

(1) Move the fixed-size window over the spectra and obtain m non-overlapping and equidistant intervals. Take the variable with the largest norm in each interval as the representative variable and denote it as $\mathbf{z} = (z_1, \ldots, z_m)$.
(2) Define the projection operator in the i -th iteration $P^i_{j_0} = I - \frac{z^i_{j_0}(z^i_{j_0})'}{(z^i_{j_0})'z^i_{j_0}}$, with the initialized starting variable $z^i_{j_0}, j_0 \in \{1, \ldots, m\}$.
(3) Compute the projected representative variable vector and take it as the updated variable in the next iteration, i.e. $z^{i+1}_j = P^i_{j_0} z^i_j, j = 1, \ldots, m$.
(4) Update the starting variable with the one that has the largest norm, i.e. $\max_{j=1,\ldots,m} \left\| z^{i+1}_j \right\|$.
(5) Run step (2)–(5) iteratively for N times, $N = 1, \ldots, (m-1)$ and obtain a vector of N starting variables.
(6) Run step (2)–(6) for different initialized starting variable $j_0 = 1, \ldots, m$ and obtain m chains of $(m-1)$ starting variable vectors.
(7) Substitute the starting variable vectors with the corresponding intervals. Establish $m(m-1)$ PLS models on the various combinations of intervals. The optimal combination of intervals is determined in terms of the model prediction performance.

It is worth noting that intervals are represented by individual variables. The representative variables are then analyzed in a similar way as SPA. The projection operator aims to reduce the collinearity among variables and find the variable that contains the largest amount of information. A chain of variable combinations is recursively searched. The optimal intervals are thus searched among the corresponding interval combinations. Empirical evidence shows that iSPA outperforms SPA by

predictive performance and exhibits a better robustness concerning the difference between the validation and test set [48]. Such conclusion enhances the necessity of selecting intervals rather than single variables in datasets with highly correlated variables.

4. Datasets and software

Three datasets, including milk dataset, tobacco dataset and soil dataset, were used to validate six of the above approaches: PLS, iPLS, MWPLS, ENPLS, SIS-iPLS, and FOSS.

4.1. Milk dataset

The milk dataset [46] is acquired directly from the local market in Changsha, China. The spectrum contains 1557 wavelength points recorded from 10000 cm^{-1} to 4000 cm^{-1} with an interval of 4 cm^{-1} . The dataset consists of 67 samples and we consider the protein of milk as the property of interest. All 67 samples were randomly split into 47 samples (70% of the dataset) for calibration and 20 samples (30% of the dataset) for test.

4.2. Tobacco dataset

The tobacco dataset [68] contains 300 samples and 1557 spectral points from 10000 cm^{-1} to 4000 cm^{-1} at 4 cm^{-1} interval. The total nicotine of the tobacco samples is employed as the response. 210 samples (70% of the dataset) were randomly sampled from all samples and used for training the model. The remaining 90 samples (30% of the dataset) were used as the independent test set.

4.3. Soil dataset

The soil dataset [69] contains 108 samples and the wavelength ranges from 400 nm to 2500 nm (visible and near infrared spectrum). In this paper, the 1100–2500 nm range of NIR is chosen and constitutes 700 spectral points according to [69]. Soil organic matter (SOM) is considered as the property of interest. The dataset was randomly divided into a calibration set containing 75 samples (70% of the dataset) and a test set containing the remaining samples.

4.4. Software

Experiments of PLS, iPLS, EN-PLS, and SIS-iPLS were carried out in R (Version 3.3.2) on a PC with Intel Core i7 2.7 GHz CPU and 32 GB RAM. PLS and iPLS models were fitted using the R package pls [70] and mdatools [71], respectively. For the EN-PLS and SIS-iPLS methods, in-house R implementations were used. The MWPLS and FOSS models were implemented and fitted in MATLAB (Version 2015a, The MathWorks, Inc.).

5. Results and discussion

5.1. Details of experiments

To illustrate performance of PLS, iPLS, MWPLS, EN-PLS, SIS-iPLS, and FOSS, three datasets were used to benchmark these algorithms. For each dataset, wavelength intensities were centered to have zero means before modeling. Calibration sets were employed for variable selection and establishment of PLS models, while independent test sets were used to evaluate calibration models. Multiple evaluation measures, such as the root mean squares error of prediction (RMSEP), root mean squares error of calibration (RMSEC) were exploited to access the model performance. Also, the optimal number of latent variables (nLV) for PLS models and the number of selected variables (nVAR) were recorded for a comprehensive view of model performances. Each method was conducted 50 times to guarantee the reproducibility and stability of experiments.

Due to the different information contained in each dataset, it is necessary to set proper and possibly different number of intervals in iPLS for varied datasets. Based on the previous work by Refs. [46,69] and our experience, the number of intervals for milk, tobacco, and soil datasets are set to 40, 40, and 18, respectively. To ensure fair comparison, we set the window width in MWPLS to be equal to that in iPLS. In the process of calibration, parameters in the elastic net and the number of latent variables were optimized by 10-fold cross validation.

5.2. Milk dataset

Results of the milk dataset are displayed in Table 1 and Fig. 1. As we can see, MWPLS showed the lowest RMSEP (0.0411), followed by FOSS (0.0436). The performance of iPLS, EN-PLS, and SIS-iPLS were not desirable with the RMSEP of 0.0457, 0.0659, and 0.0752. Compared to the full-spectrum PLS, the RMSEP values of MWPLS and FOSS decreased by 8.3% and 2.7%, respectively.

The frequency of the selected variables in 50 experiments is demonstrated in Fig. 1(b)–(f). Fig. 1(a) shows the whole spectrum. The wavelengths selected by MWPLS are similar to those selected by FOSS near the range 4500–4850 cm^{-1}, which correspond to the third overtone of C—H bending of —CH_2 group and C═O carbonyl stretch, second overtone of primary amide [46]. On the other hand, MWPLS tends to select a few extra (potentially uninformative) spectral intervals, for example, 6500–7000 cm^{-1} , which leads to more complex models than FOSS. It is worth noting that wavelengths near 4000–4040 cm^{-1} , 4320–4350 cm^{-1} and 4700–4800 cm^{-1} were frequently selected by EN-PLS. These regions are related to the second overtone of secondary amine and C═O carbonyl stretch, second overtone of primary amide. However, EN-PLS tends to select a small number of variables and thus may miss some important informative regions, which could possibly result in a relatively poor prediction performance. As for SIS-iPLS, it obtains regions around 4000–4200 cm^{-1}, 4750–4810 cm^{-1} as well as 9928–10000 cm^{-1}, which is related to the second overtone of N—H stretch [46]. But SIS-iPLS also selected some other regions, such as 4800–5300cm^{-1} and 8400–8700 cm^{-1} , which may be uninformative and cause its relatively high RMSEP value.

5.3. Tobacco dataset

Table 2 and Fig. 2 present experimental results of the tobacco dataset. FOSS and iPLS showed almost identical performance with the RMSEP of 0.0049 while FOSS has a slightly larger standard deviation. Their performance is followed by MWPLS (0.0058) and PLS (0.0067). The average RMSEP value of SIS-iPLS and EN-PLS reaches 0.0072 and 0.0081, respectively. It can be seen that the RMSEP values of iPLS, FOSS, and MWPLS are lower than the baseline method, which indicates that models constructed using the interval selection methods exhibit good predictive performances. It should be pointed out that SIS-iPLS and EN-PLS have mildly higher RMSEP values than PLS, but they selected a much smaller number of variables (139.2 and 29.1) on average. This indicates a potential improvement in the model simplicity and interpretability made by SIS-iPLS and EN-PLS, at the cost of losing some predictive performance.

Fig. 2 shows the results of the selected variables using five methods on the tobacco dataset. The wavelengths around 4389–4474 cm^{-1} were commonly selected by the five interval selection methods. This region is associated to the combination of the fundamental stretching and bending vibrations of C—H/C—C [72]. It is necessary to point out that SIS-iPLS selects not only the region 4389–4474 cm^{-1} , but also 4589–4690 cm^{-1} , which is assigned to the second overtone of N—H bending [46]. In the meantime, some additional regions were picked out as well, such as the region around 9000 cm^{-1} , which may be redundant and uninformative wavelengths and result in the undesirable RMSEP value as in Table 2. It is obvious that only a very small number of variables are selected by EN-PLS. This may increase the possibility for EN-PLS to omit some

important and informative wavelengths, thus probably limits the predictive performance of the model.

5.4. Soil SOM dataset

The results of soil SOM dataset are reported in Table 3 and Fig. 3. From Table 3, we can see that FOSS outperforms other methods with the value of RMSEP (0.3371). However, FOSS provides a less stable set of variables with the standard deviation of 108.7, compared to the other methods. The RMSEP values of MWPLS (0.3442) and SIS-iPLS (0.4600) are also lower than PLS (0.5050). As above, EN-PLS selected much fewer variables (54.8) in the total number of variables (700). This suggests that EN-PLS could miss some important spectral bands (see Fig. 4) (see Table 4).

The wavelengths selected by different methods on the soil SOM data are presented in Fig. 3. SIS-iPLS frequently obtained the region near 1920–1940 nm, 2324–2354 nm and 2440–2460 nm, which were verified to be informative spectra regions [69]. The regions selected by iPLS are similar to those selected by FOSS, except that iPLS selected some unknown spectral bands more frequently, such as 1700–1800 nm, which may result in a slightly worse RMSEP value compared with FOSS. EN-PLS managed to select some informative regions, such as 1910–1930 nm, which may "indicate O—H groups in water or various functional groups present in cellulose, lignin, glucan, starch, pectin, and humic acid." [69] In addition, the range of 2000–2030 nm were frequently selected only by EN-PLS among all other methods, so this region might be uninformative and negatively influenced the prediction performance in terms of RMSEP.

5.5. Discussion

According to our benchmarking results, the intervals selection methods, iPLS, MWPLS, EN-PLS, SIS-iPLS, and FOSS, can improve the model accuracy with varying degrees when compared to the full spectrum PLS regression. The perfect interval selection method does not exist. Efforts should be made to choose suitable methods for a given dataset.

Based on the analysis of the results, we can draw some preliminary conclusions. The two classic interval selection methods, iPLS and MWPLS, select plenty of variables, which could be associated with the value of the window width. Similarly, SIS-iPLS tends to select many variables, but still much fewer than the full spectrum of variables. A great many of variables are successfully screened out. EN-PLS selects the least number of variables in the three datasets. However, its predictive performance is not outstanding. The reason is probably that the selected variables lost too much useful information of the original space, thus weaken the model's ability to explain the response variable. However, this does not indicate that models with more variables are better than those with fewer variables. The additional variables selected by SIS-iPLS did not contribute to better predictive performance, and this may come from the inclusion of some unimportant variables. In the three experiments, FOSS shows a remarkable performance among all methods, while its computational complexity is unsatisfactory compared to other methods.

6. Summary

In this paper, we focused on and reviewed five classes of interval selection methods: classic methods, penalty-based, sampling-based, correlation-based, and projection-based methods. Classic methods predetermine the partition of the spectrum. Thus, many methods, such as PLS and group lasso, can be plugged-in for calibration. Consequently, the classic methods are flexible and can work with different regression techniques. However, the construction of the intervals is subjective rather than data-driven and could easily fail to include the information of the response variable, which can be further investigated [73]. Besides, assessing the predictive performance of a mass of intervals burdens the computation. In contrast, the penalty-based methods take the response variable into consideration and construct the intervals adaptively based on the properties of the penalty. Nevertheless, since the parameters usually have a heavy influence on the penalty, the proper tuning of these parameters can be critical. The improperly tuned parameters can result in models with bad overall performance. Choosing appropriate values for the tuning parameters can be tough and compromise the computational time. The sampling-based methods also enjoy the feature of adaptively constructed intervals. Although methods based on sampling can produce models with high performance, the fluctuation of the model performance can be a major limitation, not to mention the difficulty of achieving reproducible and consistent conclusions. This is primarily caused by the uncertainty derived from the sampling procedure, specifically, the properties of the population, the way of sampling and the methods for estimation [62]. The correlation-based methods have an outstanding advantage for being fast, easy to compute as well as being scalable. These features make such methods more suitable for large-scale data than methods in other categories in light of the computation. On the other hand, the correlation criteria employed to rank the variables can reflect the linear relevance between the variables and the response but could fail to explore the potential nonlinear relationships. Additionally, the determination of the threshold is also a major challenge. The projection-based methods employ the projection operator to ensure the inclusion of important variables. Moreover, the recursive search for various combinations of variables increases the possibility of hitting the optimal variable subset. Unfortunately, the exhaustive search is precisely the main cause of their high computational cost.

The interval selection methods reviewed in this paper are applicable in the case where the number of variables is much larger than the number of samples, and high correlations exist among the variables. The spectroscopic data investigated in this paper is a representative example of such type of data. To provide a comprehensive understanding and profound insights into different methods, three real datasets were employed to evaluate the model performances. There are no such perfect methods but only proper methods for particular datasets. The scope of the methods and experiments is limited in this review, but we hope it can offer some general and informative guidelines for spectroscopic data modeling.

Acknowledgement

We thank the editor and the referee for constructive suggestions that substantially improved this work. This work is financially supported by the National Natural Science Foundation of China (Grant No. 11271374), the Key Laboratory for Mixed and Missing Data Statistics of the Education Department of Guangxi Province (Grant No. GXMMSL201404), and the Mathematics and Interdisciplinary Sciences Project, and the Innovation Program of Central South University.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.chemolab.2017.11.008.

References

[1] T. Hasegawa, Principal component regression and partial least squares modeling, in: J.M. Chalmers, P.R. Griffiths (Eds.), Handbook of Vibrational Spectroscopy, John Wiley & Sons, New York, 2002, pp. 2293–2312.
[2] I.M. Johnstone, D.M. Titterington, Statistical challenges of high-dimensional data, Philos. Trans. A. Math. Phys. Eng. Sci. 367 (2009) 4237–4253.
[3] J. Fan, R. Li, Statistical challenges with high dimensionality: feature selection in knowledge discovery, in: Proceedings of the International Congress of Mathematicians, 2006, pp. 595–622.
[4] J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space, J. R. Stat. Soc. B 70 (2008) 849–911.
[5] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Anal. Chim. Acta 185 (1986) 1–17.
[6] R. Rosipal, N. Krämer, Overview and recent advances in partial least squares, Lect. Notes. Comput. Sci. 3940 (2006) 34–51.
[7] Q.S. Xu, S. de Jong, P. Lewi, D.L. Massart, Partial least squares regression with Curds and Whey, Chemom. Intell. Lab. Syst. 71 (2004) 21–31.

[8] Y.W. Lin, B.C. Deng, Q.S. Xu, Y.H. Yun, Y.Z. Liang, The equivalence of partial least squares and principal component regression in the sufficient dimension reduction framework, Chemom. Intell. Lab. Syst. 150 (2016) 58–64.
[9] D.V. Nguyen, D.M. Rocke, Tumor classification by partial least squares using microarray gene expression data, Bioinformatics 18 (2002) 39–50.
[10] J. Nilsson, S. de Jong, A.K. Smilde, Multiway calibration in 3D QSAR, J. Chemom. 11 (1997) 511–524.
[11] J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space, Stat. Sin. 20 (2010) 101–148.
[12] X.B. Zou, J.W. Zhao, M.J.W. Povey, M. Holmes, H.P. Mao, Variables selection methods in near-infrared spectroscopy, Anal. Chim. Acta 667 (2010) 14–32.
[13] A. Höskuldsson, Variable and subset selection in PLS regression, Chemom. Intell. Lab. Syst. 55 (2001) 23–38.
[14] E.V. Thomas, A primer on multivariate calibration, Anal. Chem. 66 (1994) 795–804.
[15] F.G. Blanchet, P. Legendre, D. Borcard, Forward selection of spatial explanatory variables, Ecology 89 (2008) 2623–2632.
[16] J.M. Sutter, J.H. Kalivas, Comparison of forward selection, backward elimination, and generalized simulated annealing for variable selection, Microchem. J. 47 (1993) 60–66.
[17] S. Derksen, H.J. Keselman, Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables, Br. J. Math. Stat. Psychol. 45 (1992) 265–282.
[18] I.E. Frank, Intermediate least squares regression method, Chemom. Intell. Lab. Syst. 1 (1987) 233–242.
[19] A.G. Frenich, D. Jouan-Rimbaud, D.L. Massart, S. Kuttatharmmakul, M.M. Galera, J.L.M. Vidal, Wavelength selection method for multicomponent spectrophotometric determinations using partial least squares, Analyst 120 (1995) 2787–2792.
[20] S. Wold, E. Johansson, M. Cocchi, PLS-partial least squares projections to latent structures, in: H. Kubinyi (Ed.), 3D QSAR in Drug Design, Theory, Methods, and Applications, ESCOM Science Publishers, Leiden, 1993, pp. 523–550.
[21] R. Tibshirani, Regression selection and shrinkage via the lasso, J. R. Stat. Soc. B 58 (1996) 267–288.
[22] J. Fan, Comments on "Wavelets in statistics: a review" by A. Antoniadis, J. Ital. Stat. Soc. 6 (1997) 131–138.
[23] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, J. Am. Stat. Assoc. 96 (2001) 1348–1360.
[24] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, J. R. Stat. Soc. B 72 (2010) 3–25.
[25] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, Model population analysis for variable selection, J. Chemom. 24 (2010) 418–423.
[26] H.D. Li, Q.S. Xu, Y.Z. Liang, Random frog: an efficient reversible jump Markov Chain Monte Carlo-like approach for variable selection with applications to gene selection and disease classification, Anal. Chim. Acta 740 (2012) 20–26.
[27] Y.H. Yun, W.T. Wang, M.L. Tan, Y.Z. Liang, H.D. Li, D.S. Cao, H.M. Lu, Q.S. Xu, A strategy that iteratively retains informative variables for selecting optimal variable subset in multivariate calibration, Anal. Chim. Acta. 807 (2014) 36–43.
[28] B.C. Deng, Y.H. Yun, Y.Z. Liang, L.Z. Yi, A novel variable selection approach that iteratively optimizes variable space using weighted binary matrix sampling, Analyst 139 (2014) 4836–4845.
[29] B.C. Deng, Y.H. Yun, D.S. Cao, Y.L. Yin, W.T. Wang, H.M. Lu, Q.Y. Luo, Y.Z. Liang, A bootstrapping soft shrinkage approach for variable selection in chemical modeling, Anal. Chim. Acta 908 (2016) 63–74.
[30] S. Kirkpatrick, C.D. Gelatt, M.P. Vecchi, Optimization by simulated annealing, Science 220 (1983) 671–680.
[31] A. Verikas, M. Bacauskiene, Feature selection with neural networks, Pattern Recognit. Lett. 23 (2002) 1323–1335.
[32] R. Leardi, Genetic algorithms in chemometrics and chemistry: a review, J. Chemom. 15 (2001) 559–569.
[33] M.C.U. Araújo, T.C.B. Saldanha, R.K.H. Galvão, T. Yoneyama, H.C. Chame, V. Visani, The successive projections algorithm for variable selection in spectroscopic multicomponent analysis, Chemom. Intell. Lab. Syst. 57 (2001) 65–73.
[34] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, C. Sterna, Elimination of uninformative variables for multivariate calibration, Anal. Chem. 68 (1996) 3851–3858.
[35] S. Ye, D. Wang, S. Min, Successive projections algorithm combined with uninformative variable elimination for spectral variable selection, Chemom. Intell. Lab. Syst. 91 (2008) 194–199.
[36] Y.W. Lin, N. Xiao, L.L. Wang, C.Q. Li, Q.S. Xu, Ordered homogeneity pursuit lasso for group variable selection with applications to spectroscopic data, Chemom. Intell. Lab. Syst. 168 (2017) 62–71.
[37] B.C. Deng, Y.H. Yun, P. Ma, C.C. Lin, D.B. Ren, Y.Z. Liang, A new method for wavelength interval selection that intelligently optimizes the locations, widths and combinations of the intervals, Analyst 140 (2015) 1876–1885.
[38] L. Norgaard, 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. 54 (2000) 413–419.
[39] J.H. Jiang, R.J. Berry, H.W. Siesler, Y. Ozaki, Wavelength interval selection in multicomponent spectral analysis by moving window partial least-squares regression with applications to mid-infrared and near-infrared spectroscopic data, Anal. Chem. 74 (2002) 3555–3565.
[40] G.H. Fu, Q.S. Xu, H.D. Li, D.S. Cao, Y.Z. Liang, Elastic net grouping variable selection combined with partial least squares regression (EN-PLSR) for the analysis of strongly multi-collinear spectroscopic data, Appl. Spectrosc. 65 (2011) 402–408.
[41] X. Huang, Y.P. Luo, Q.S. Xu, Y.Z. Liang, Elastic net wavelength interval selection based on iterative rank PLS regression coefficient screening, Anal. Methods 9 (2017) 672–679.
[42] B. Liquet, P.L. Micheaux, B.P. Hejblum, R. Thiébaut, Group and sparse group partial least square approaches applied in genomics context, Bioinformatics 32 (2015) 35–42.
[43] L.P. Brás, M. Lopes, A.P. Ferreira, J.C. Menezes, A bootstrap-based strategy for spectral interval selection in PLS regression, J. Chemom. 22 (2008) 695–700.
[44] R. Gosselin, D. Rodrigue, C. Duchesne, A Bootstrap-VIP approach for selecting wavelength intervals in spectral imaging applications, Chemom. Intell. Lab. Syst. 100 (2010) 12–21.
[45] Y.W. Lin, B.C. Deng, L.L. Wang, Q.S. Xu, L. Liu, Y.Z. Liang, Fisher optimal subspace shrinkage for block variable selection with applications to NIR spectroscopic analysis, Chemom. Intell. Lab. Syst. 159 (2016) 196–204.
[46] Y.H. Yun, H.D. Li, L.R. Leslie, W. Fan, J.J. Wang, D.S. Cao, Q.S. Xu, Y.Z. Liang, An efficient method of wavelength interval selection based on random frog for multivariate spectral calibration, Spectrochim. Acta. Mol. Biomol. Spectrosc. 111 (2013) 31–36.
[47] J. Xu, Q.S. Xu, C.O. Chan, D.K. Mok, L.Z. Yi, F.T. Chau, Identifying bioactive components in natural products through chromatographic fingerprint, Anal. Chim. Acta 870 (2015) 45–55.
[48] A.A. Gomes, R.K.H. Galvão, M.C.U. Araújo, G. Véras, E.C. Silva, The successive projections algorithm for interval selection in PLS, Microchem. J. 110 (2013) 202–208.
[49] P. Geladi, Notes on the history and nature of partial least squares (PLS) modelling, J. Chemom. 2 (1988) 231–246.
[50] A. Höskuldsson, PLS regression methods, J. Chemom. 2 (1988) 211–228.
[51] P.D. Sampson, A.P. Streissguth, H.M. Barr, F.L. Bookstein, Neurobehavioral effects of prenatal alcohol: Part II. Partial least squares analysis, Neurotoxicol. Teratol. 11 (1989) 477–491.
[52] S. de Jong, SIMPLS: an alternative approach to partial least squares regression, Chemom. Intell. Lab. Syst. 18 (1993) 251–263.
[53] Q.S. Xu, Y.Z. Liang, H.L. Shen, Generalized PLS regression, J. Chemom. 15 (2001) 135–148.
[54] L. Munck, J.P. Nielsen, B. Møller, S. Jacobsen, I. Søndergaard, S.B. Engelsen, L. Nørgaard, R. Bro, Exploring the phenotypic expression of a regulatory proteome-altering gene by spectroscopy and chemometrics, Anal. Chim. Acta 446 (2001) 169–184.
[55] X.B. Zou, J.W. Zhao, Y.X. Li, Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of "Fuji" apple based on BiPLS and FiPLS models, Vib. Spectrosc. 44 (2007) 220–227.
[56] R. Leardi, L. Nørgaard, Sequential application of backward interval partial least squares and genetic algorithms for the selection of relevant spectral regions, J. Chemom. 18 (2004) 486–497.
[57] Y.P. Du, Y.Z. Liang, J.H. Jiang, R.J. Berry, Y. Ozaki, Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares, Anal. Chim. Acta 501 (2004) 183–191.
[58] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. B 67 (2005) 301–320.
[59] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. B 68 (2006) 49–67.
[60] K.A. le Cao, D. Rossouw, C. Robert-Granié, P. Besse, A sparse PLS for variable selection when integrating omics data, Stat. Appl. Genet. Mol. Biol. 7 (2008) 1–29.
[61] H. Chun, S. Keleş, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, J. R. Stat. Soc. B 72 (2010) 3–25.
[62] J.F. Wang, A. Stein, B.B. Gao, Y. Ge, A review of spatial sampling, Spat. Stat 2 (2012) 1–14.
[63] B. Efron, R.J. Tibshirani, An Introduction to the Bootstrap, Chapman & Hall, Boca Raton, 1993.
[64] A. Lazraq, R. Cléroux, J.P. Gauchi, Selecting both latent and explanatory variables in the PLS1 regression model, Chemom. Intell. Lab. Syst. 66 (2003) 117–126.
[65] P. Hall, J.L. Horowitz, B.Y. Jing, On blocking rules for the bootstrap with dependent data, Biometrika 82 (1995) 561–574.
[66] W.D. Fisher, On grouping for maximum homogeneity, J. Am. Stat. Assoc. 53 (1958) 789–798.
[67] W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications, Biometrika 57 (1970) 97–109.
[68] K. Zheng, Q. Li, J. Wang, J. Geng, P. Cao, T. Sui, X. Wang, Y. Du, Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra, Chemom. Intell. Lab. Syst. 112 (2012) 48–54.
[69] R. Rinnan, A. Rinnan, Application of near infrared reflectance (NIR) and fluorescence spectroscopy to analysis of microbiological and chemical properties of arctic soil, Soil Biol. Biochem. 39 (2007) 1664–1673.
[70] B.H. Mevik, R. Wehrens, The pls package: principal component and partial least squares regression in R, J. Stat. Softw. 18 (2007) 1–24.
[71] S. Kucheryavskiy, Mdatools: Multivariate Data Analysis for Chemometrics, 2017. https://cran.r-project.org/package=mdatools. (Accessed 30 January 2017).
[72] H. Xu, Z. Liu, W. Cai, X. Shao, A wavelength selection method based on randomization test for near-infrared spectral analysis, Chemom. Intell. Lab. Syst. 97 (2009) 189–193.
[73] J.F. Wang, T.L. Zhang, B.J. Fu, A measure of spatial stratified heterogeneity, Ecol. Indic. 67 (2016) 250–256.

View publication stats