Chemometrics and Intelligent Laboratory Systems 168 (2017) 62–71

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

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

Ordered homogeneity pursuit lasso for group variable selection with applications to spectroscopic data

You-Wu Lin [a,b], Nan Xiao [b,c], Li-Li Wang [b], Chuan-Quan Li [b], Qing-Song Xu [b,*]

[a] School of Mathematics and Statistics, Guangxi Teachers Education University, Nanning 530023, PR China
[b] School of Mathematics and Statistics, Central South University, Changsha 410083, PR China
[c] Seven Bridges Genomics, 1 Main Street, Cambridge, MA 02142, USA

ARTICLE INFO

Keywords:
Lasso
Homogeneity pursuit
Sparse learning
Variable ordering
Grouping effect
Partial least squares

ABSTRACT

In high-dimensional data modeling, variable selection methods have been a popular choice to improve the prediction accuracy by effectively selecting the subset of informative variables, and such methods can enhance the model interpretability with sparse representation. In this study, we propose a novel group variable selection method named ordered homogeneity pursuit lasso (OHPL) that takes the homogeneity structure in high-dimensional data into account. OHPL is particularly useful in high-dimensional datasets with strongly correlated variables. We illustrate the approach using three real-world spectroscopic datasets and compare it with four state-of-the-art variable selection methods. The benchmark results on real-world data show that the proposed method is capable of identifying a small number of influential groups and has better prediction performance than its competitors. The OHPL method and the spectroscopic datasets are implemented and included in an R package OHPL available from https://ohpl.io.

1. Introduction

High-dimensional data analysis problems arise from many frontiers of scientific disciplines and technological revolutions [1,2]. For example, the problems of high-dimensional data have been produced in the context of chemometrics [3]. In the molecular descriptors datasets or spectral datasets, several hundreds of expressions of molecules or wavelengths are potential variables [4–6]. In biomedical studies, enormous numbers of magnetic resonance spectroscopy data or magnetic resonance images (MRI) and functional MRI data with tens of hundreds of features are available [7,8]. In the field of bioinformatics, the number of variables reaches thousands or even more [9]. The validity of high-dimensional data along with new scientific problems creates opportunities and poses challenges for the development of new statistical techniques. Variable selection [10–13] and dimensionality reduction [14,15] play the pivotal role in almost all modern statistical research and discoveries driven by high-dimensional data.

The topic of variable selection has a long history, and various techniques and methods have been developed. One important class of methods is summarized with the term penalized methods or regularized methods. There have been a number of works in statistics and machine learning dealing with penalization in a broad spectrum of problems. A common feature of classical model selection criteria, such as C_p, AIC, and BIC, is a combinatorial optimization problem, which is NP-hard. Since the computational time of traditional penalized variable selection tools increases exponentially with the data dimensionality, they become inadequate or completely fail under modern high-dimensional settings. To overcome this difficulty, many kinds of new procedures and methods have been developed. For example, L_2-penalty (ridge regression [16]) is commonly used among statisticians. The drawback of ridge regression is that it cannot provide a sparse model, for it always keeps all the variables in the model, although the coefficients of many variables could be near zero. Followed by L_γ-penalty (bridge regression [17]) which was first considered as a unifying framework to understand penalized regression and variable selection. L_1-penalty, a special case of bridge regression named lasso, was introduced by Tibshirani [11]. Owing to the nature of the L_1-penalty, the lasso does not only offer continuous shrinkage but also performs automatic variable selection to produce a sparse model. Variable selection methods based on L_1-penalty have attracted plenty of research efforts due to its sparsity-inducing property, convenient convexity, and excellent theoretical guarantees. Various generalizations and variants of the lasso were developed, such as elastic net (EN) [18], fused lasso (Flasso) [19], and grouped lasso [20], to name a few. These methods have achieved great success in diverse fields of sciences, such as

* Corresponding author.
E-mail address: qsxu@csu.edu.cn (Q.-S. Xu).

http://dx.doi.org/10.1016/j.chemolab.2017.07.004
Received 23 January 2017; Received in revised form 8 July 2017; Accepted 10 July 2017
Available online 13 July 2017

genomics, bioinformatics, econometrics, and finance, to address a variety of problems. Recently, regularization methods have received increasing attention due to their sparse representations and high prediction accuracy in chemometrics. For example, Filzmoser and others provided a comparison between sparse methods and non-sparse counterparts to analyze several high-dimensional datasets from chemometrics and showed that the sparseness in the model could lead to an improvement of the prediction or classification performance [21]. Kalivas used L_1-penalty, L_2-penalty, and their combined forms to full wavelength or sparse spectral multivariate calibration models or maintenance [22]. Shahbazikhah, Kalivas, and others applied the L_1-penalty to select basis set vectors for multivariate calibration and calibration updating [23]. Randolph used adaptive penalties on generalized Tikhonov regularization to build regression models for spectroscopy data [24]. Higashi and others applied sparse regression methods to select fluorescence wavelengths for accurate prediction of food properties [25].

As described above, dimensionality reduction is another effective approach to solving the high-dimensional problems in modern statistical research and scientific discoveries. Partial least squares (PLS) and principal component regression (PCR) are two powerful and popular methods for compressing high-dimensional data sets. Lin et al. proved the equivalence of PLS and PCR under the sufficient dimension reduction (SDR) framework [26]. PLS is particularly useful when multicollinearity exists among the variables, and the number of variables (p) is much larger than the number of samples (n). Since the method has good predictive performance, it has been applied to different fields. For more details, see the recent review in Mehmood and Ahmed [27] and the references therein. In addition, many researchers have either theoretically or experimentally proved that additional variable selection could further improve the prediction accuracy of PLS models. Many variable selection criteria and sparsity-inducing procedures based on the PLS model have been proposed in the literature. For example, Chun and Keles proposed sparse partial least squares (SPLS) regression for simultaneous dimension reduction and variable selection with applications to the analysis of gene expression data [9]. Chung and Keles also introduced sparse partial least squares classification for high-dimensional data [28]. Filzmoser and others provided a comparison between PLS and SPLS on analyzing several NIR spectroscopic datasets and showed that the SPLS method outperforms original PLS [21]. In a typical spectroscopic analysis, there are often two types of variables selection or wavelength selection methods. One is to select individual wavelengths: select some wavelengths discretely and analyze associations between the property of interest and individual wavelengths [29–32]. The other is to determine wavelength intervals (groups): study the associations between the property of interest and intervals of wavelengths [33–36]. As was precisely summarized in Ref. [22]: "Forming models with wavelengths selected can be thought of as forming sparse models."

Although discrete variable selection (individual wavelengths) methods such as the lasso have achieved great success in many applications, they do not take the existing data structure into consideration. The underlying structural information in the data may improve the regression or classification performance significantly, and help identify the important variables. For example, in spectroscopic data, there are two characteristics of the variables (wavelengths): one is the natural spatial order of wavelengths, and the other one is that consecutive variables carry similar information. Based on these characteristics of data, Lin and others proposed a group variable selection method called Fisher optimal subspace shrinkage (FOSS) with applications to NIR spectroscopic data. The intuition behind the method is that the regression coefficients of consecutive wavelengths should have close values. Empirical results showed that the performance of their method outperforms its competitors [37].

Recently, Ke, Fan, and Wu considered a more general concept than sparsity: homogeneity. The intuitive goal of exploring homogeneity is to divide the regression coefficients into several groups such that the values of regression coefficients in the same group are the same or close and the values of regression coefficients in different groups are significantly different to each other. Sparsity is a special case of homogeneity where a large number of groups have zero coefficients solely. For example, homogeneity exists when the slope parameters in the regression model come from a network of dependent genes or correspond to neighboring geographical regions. Besides enhancing predictive performance, the successful detection of homogeneity can also enable the regression model to capture the "natural" structure of the data. Several approaches have been proposed to detect the homogeneity in parameters, for example, Shen and Huang [39], Ke, Fan, and Wu [38], Ke, Li, and Zhang [40]. However, the methods in these works are confined to genomics data or panel data, where the datasets do not necessarily have strong correlations between consecutive predictors. The settings addressed by these researchers are different from what we investigate in this paper. In this study, we propose a novel method named ordered homogeneity pursuit lasso (OHPL) to explore homogeneity in spectroscopic data. The main idea of OHPL is to use the homogeneity of regression coefficients to construct groups or wavelength intervals. Then, the group prototype (the most informative variable in each group) is extracted and the sparse regression techniques, such as lasso, are applied to these prototypes. This step criterion is inspired by Refs. [41,42]. Finally, the selected prototypes, as well as their corresponding groups are used to build a PLS model. OHPL has two advantages compared with the original lasso. First, it can select more than n variables (n is the sample size). Second, it can identify the homogenous grouping effect that naturally represents the spatial structure in the predictors. Some of the state-of-the-art penalized regression methods, including lasso, elastic net, fused lasso, and Sparse PLS are used for comparison. The empirical results show that, when compared with the PLS model, the OHPL model has a lower prediction error and better interpretability. Comparing with lasso regressions, OHPL models lead to improved performance on prediction and encourages the grouping effect. Comparing with the other traditional group variable selection methods, such as elastic net and fused lasso, the OHPL method leads to higher predictive accuracy and better detection of relevant variables.

This paper is organized as follows. Section 2 briefly outlines the four penalized regression methods and describes the proposed method. A brief introduction and summary of three near-infrared spectroscopy datasets are given in Section 3. Section 4 presents and discusses the benchmark results on the three datasets. Finally, Section 5 summarizes the proposed method and concludes this paper. The appendix gives the details of Fisher optimal partitions algorithm.

2. Theory and algorithm

In this paper, we consider the following high-dimensional linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{1}$$

where $\mathbf{X}$ is an $n \times p$ design matrix and assume $\mathbf{x}_1, \cdots, \mathbf{x}_p$ are standardized, $\mathbf{y}$ is an $n \times 1$ vector of response and assume it is centered, $\boldsymbol{\beta}$ is a vector of parameters, $\boldsymbol{\varepsilon}$ is an $n \times 1$ error random vector with mean zero ($\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$) and variance ($\mathrm{var}(\varepsilon) = \sigma^2\mathbf{I}$).

2.1. Lasso

The lasso is one of the most popular penalized regression techniques, which essentially imposes a constraint on the L_1 norm of the regression coefficients [11]. The lasso estimator is defined as

$$\widehat{\boldsymbol{\beta}}_{lasso} = \underset{\boldsymbol{\beta}\in\mathbf{R}^p}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y}\text{-}\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 \tag{2}$$

where $\lambda_1 \geq 0$ is a fixed tuning parameter, controlling the degree of sparsity in the estimate $\widehat{\boldsymbol{\beta}}_{lasso}$. In this work, we use 5-fold cross-validation

to guide the choice of the optimal tuning parameter λ_1, which is the one giving the smallest CV error. The lasso continuously shrinks regression coefficients toward zero, thus improving its prediction ability via the bias-variance trade-off. Furthermore, because of the properties of the L_1-penalty, several coefficients in the resulting vector $\widehat{\boldsymbol{\beta}}_{lasso}$ will be exactly zero if λ_1 is large enough. Therefore, the lasso can be regarded as a variable selection technique. It has been widely used in genomics, bioinformatics, econometrics, and finance among others. Recently, the lasso has been used for wavelength selection and was found to produce lower prediction errors than full wavelength models [22].

2.2. Elastic net (EN)

The lasso has proven to be a successful variable selection method in many modern applications, but it has some drawbacks under certain settings. First, the maximum number of non-zero coefficients estimated by lasso is limited by the number of samples [18]. Also, if a group of relevant variables is highly correlated, the lasso tends to include only one variable from the group and ignores which one is selected. To overcome these limitations, Zou and Hastie [18] developed a modified version of lasso named elastic net. The estimator of the elastic net is given as follows

$$\widehat{\boldsymbol{\beta}}_{EN} = \underset{\boldsymbol{\beta}\in\mathbf{R}^{\mathbf{p}}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2 \tag{3}$$

where $\lambda_1 \geq 0, \lambda_2 \geq 0$ are two tuning parameters. The first tuning parameter λ_1 encourages sparsity in the regression coefficients; the second tuning parameter λ_2 encourages the grouping effect. Obviously, elastic net is equivalent to the lasso penalty when $\lambda_2 = 0$, and equivalent to the ridge penalty when $\lambda_1 = 0$. It is worth noting that the parameter λ_1 and λ_2 could be transformed to an equivalent form λ and $\alpha \in [0, 1]$, where α is a weighting parameter between the ridge regression and lasso [51]. The two tuning parameters λ and α are chosen by cross-validation as suggested by Zou and Hasite [18]. We first produce a grid of α values from 0 to 1 by increment of 0.05. Then, for each fixed α, the optimal value of λ is chosen by 5-fold cross-validation. The optimal λ is the one giving the highest CV prediction accuracy.

Elastic net has been a favorable variable selection procedure for high-dimensional regression analysis of biological datasets [43] and spectroscopic datasets [22,44].

2.3. Fused lasso

The fused lasso is another modified version of lasso and is defined as

$$\widehat{\boldsymbol{\beta}}_{Flasso} = \underset{\boldsymbol{\beta}\in\mathbf{R}^{\mathbf{p}}}{\operatorname{argmin}} \frac{1}{2}\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda_1\sum_{j=1}^{p}|\beta_j| + \lambda_2\sum_{j=2}^{p}|\beta_j - \beta_{j-1}| \tag{4}$$

where $\lambda_1 \geq 0, \lambda_2 \geq 0$ are two tuning parameters. The first tuning parameter λ_1 encourages sparsity in the regression coefficients; the second tuning parameter λ_2 encourages sparsity in their differences, i.e. the smoothness of the coefficient profiles β_j as a function of j. As discussed above, the tuning parameter λ_1 and λ_2 could be rewritten to an equivalent form λ and $\gamma \in [0, 1]$, where γ is a trade-off parameter between the second term and the third term of equation (4). Here, we use 5-fold CV to select the two tuning parameters λ and γ. We first produce a grid of γ values from 0 to 1 by increment of 0.05. Then, for each fixed γ, the optimal value of λ is chosen by 5-fold cross-validation. The optimum λ and γ are the ones giving the highest CV prediction accuracy. The fused lasso is a high-performance method for exploring homogeneity in the cases where variables have certain kinds of natural ordering [19].

Fused lasso has been successfully applied to some comparative genomic hybridization data, mass spectroscopy data [19], and HIV surveillance cohort data [45].

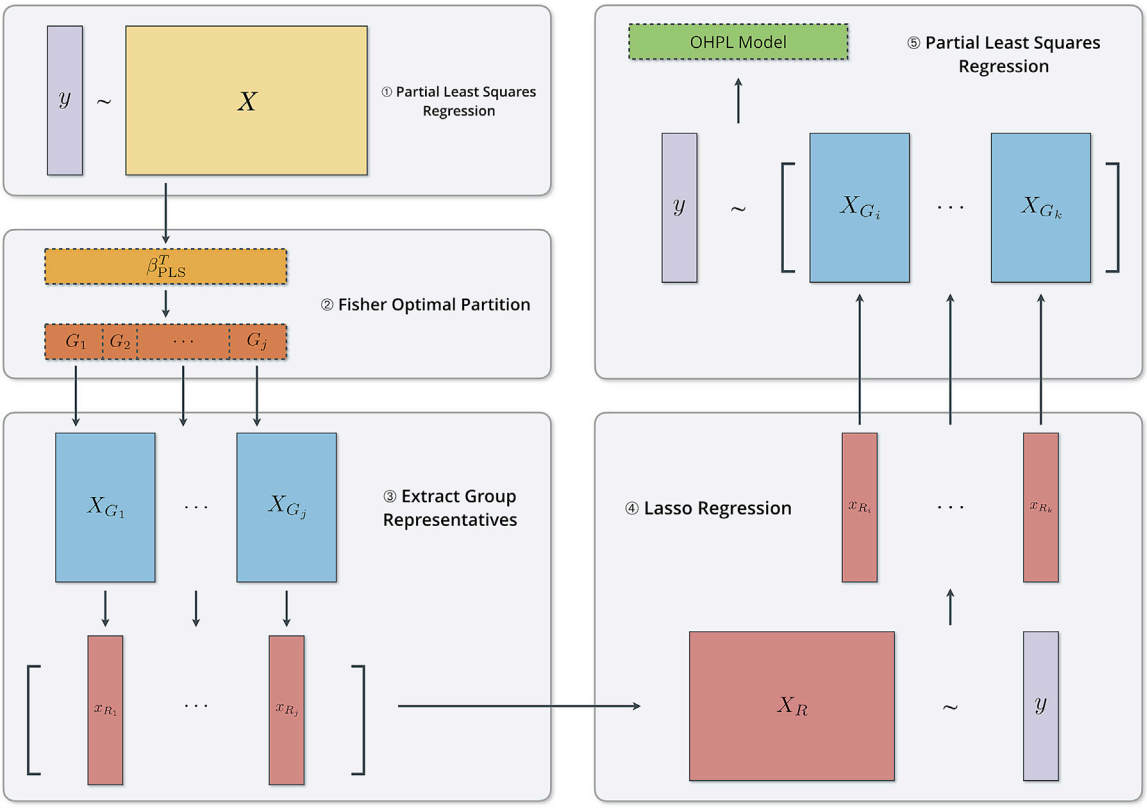

Fig. 1. The flowchart of the OHPL algorithm.

2.4. Sparse partial least squares (SPLS)

Sparse PLS is a simultaneous dimension reduction and variable selection method [9]. Its estimator is given as follows

$$\widehat{\boldsymbol{\omega}} = \underset{\mathbf{c},\boldsymbol{\omega},\|\boldsymbol{\omega}\|_2=1}{\operatorname{argmin}} \Big\{ -k\boldsymbol{\omega}^T\mathbf{M}\boldsymbol{\omega} + (1-k)(\mathbf{c}-\boldsymbol{\omega})^T\mathbf{M}(\mathbf{c}-\boldsymbol{\omega}) + \lambda_1\|\mathbf{c}\|_1 + \lambda_2\|\mathbf{c}\|_2^2 \Big\} \tag{5}$$

where $\boldsymbol{\omega}$ is the original direction vector, $\mathbf{M} = \mathbf{X}^T\mathbf{y}\mathbf{y}^T\mathbf{X}$, and $\mathbf{c}$ is a surrogate of the direction vector. SPLS has four tuning parameters $K > 0, 0 < k < 1, \lambda_1 \geq 0, \lambda_2 \geq 0$, K is the number of components, k is a parameter that controls the compromise between the first term $\boldsymbol{\omega}^T\mathbf{M}\boldsymbol{\omega}$ and the second term $(\mathbf{c}-\boldsymbol{\omega})^T\mathbf{M}(\mathbf{c}-\boldsymbol{\omega})$, λ_1 encourages sparsity in a surrogate of vector $\mathbf{c}$. We note that equation (5) imposed the L_1-penalty onto the surrogate of the weight vector $\mathbf{c}$ instead of the original weight vector $\boldsymbol{\omega}$. As was pointed out by Chun and Keles [9], this operation can produce much sparer solutions than that obtained by simply adding a penalty to the weight vector $\boldsymbol{\omega}$. In addition, they also pointed out that there are only two key tuning parameters K and λ_1 to be selected for univariate y in equation (5). Therefore, we use 5-fold CV to select two tuning parameters K and λ_1. We first produce a sequence of K values from 1 to a positive integer (e.g. 10) by increment of 1. Then, for each fixed K, the optimal value of λ_1 is chosen by 5-fold cross-validation. The optimum λ_1 and K are the ones giving the smallest CV error.

2.5. Ordered homogeneity pursuit lasso (OHPL)

The OHPL method is proposed to select informative variables with high correlation. As described above, this method uses the homogeneity of regression coefficients to construct the groups at first. Then, the group prototypes (the most correlative variable with the response y in each group) are extracted, and a sparse estimation procedure, such as lasso, is applied to these representatives. Finally, the selected prototypes, as well as their corresponding groups, are used to build a PLS model. Note that there are some parameters to be tuned in this algorithm; we will discuss how to choose them in section 2.6. The detailed procedure is described as follows.

Step 1. With the calibration set, a PLS model is built on $\mathbf{X}$, and the regression coefficients $\boldsymbol{\beta}_{pls}$ are calculated.
Step 2. To construct the groups, Fisher optimal partitions algorithm [37,46] is applied to the PLS regression coefficients $\boldsymbol{\beta}_{pls}$.
Step 3. Group representatives (prototypes) extraction, one from each group, by calculating the maximum of all inner products between each group member $\mathbf{x}_i$ and the response $\mathbf{y}$ [47]. The group representatives are calculated as follows:

$$\mathbf{x}_{R_j} = \underset{i\in G_j}{\operatorname{argmax}} \left|\mathbf{x}_i^T\mathbf{y}\right| \tag{6}$$

where $\mathbf{x}_i$ is standardized, the response $\mathbf{y}$ is centered and G_j is the j-th group.

Step 4. The sparse estimation method (lasso) is applied to these representatives.
Step 5. With the selected (by the lasso) representatives according to step 4, as well as their corresponding groups, a PLS model is fitted and applied to predict the spectra of an independent test set.

The flowchart of the OHPL algorithm is showed in Fig. 1.

2.6. Parameter setting/tuning

The OHPL algorithm has three tuning parameters: the number of component K in step 1, the number of groups g in step 2, and the L_1-penalty parameter λ_1 in step 4. As described above, we use 5-fold CV to choose the value K, g, and λ_1. First, a 5-fold CV was used to optimize the component number parameter K. Then, for the fixed value K, we use grid search with 5-fold CV to optimize the number of groups g and the regularized parameter λ_1.

2.7. Performance evaluation

Recently, some new validation methods are introduced for assessing the predictive ability of the models [48]. The most efficient evaluation strategy is performed by dividing the datasets into a calibration set and an independent test set. The calibration set is used for the calibration of its parameter values, variable selection and modeling, and the independent test set is used for evaluation of the calibration model. In this study, the training set is used for developing the model, choosing tuning parameters and implementing variable selection. The independent test set is then used to assess the calibration model. In addition, the root mean square error (RMSE), coupled with the coefficient of determination (Q^2) for both the training and test sets, are used to assess the performance of the different methods. RMSE is given as follows:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\widehat{y}-y)^2}{n}} \tag{7}$$

where $\widehat{y}$ is the predicted response value and n is the number of samples.

The coefficient of determination is defined by the formula:

$$Q^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\widehat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\overline{y}_i)^2} \tag{8}$$

where $\overline{y}$ is the average of predicted value of response.

RMSEC represents the RMSE value from the training set. The notation RMSEP indicates that the RMSE is calculated from an independent test set. Q_C^2 and Q_P^2 are the coefficients of determination for these situations, respectively.

3. Datasets and software implementation

3.1. Beer dataset

The beer spectral dataset contains 60 samples published in Ref. [49]. They are recorded with a 30 mm quartz cell directly on the undiluted degassed beer and measured from 1100 to 2250 nm (576 data points) in steps of 2 nm. Original extract concentration which illustrates the substrate potential for the yeast to ferment alcohol is considered as the property of interest. Also, we randomly split the data set into a training data set containing 70% of the data and a test data set with the remaining samples.

3.2. Wheat dataset

The entire data consists of 100 wheat samples with specified protein and moisture content. Samples were measured by diffuse reflectance as log (I/R) from 1100 to 2500 nm (701 data points) in 2 nm intervals. The protein of wheat is used as dependent variable y in this study. More details are described in Ref. [50]. Besides, the dataset is split into a calibration set (70% of the samples, 70 samples) and an independent test set (30% of the samples, 30 samples) with the Monte Carlo method.

3.3. Soil dataset

The soil dataset contains 108 sample measurements from the wavelength range of 400–2500 nm (visible and near infrared spectrum),

Table 1
The benchmark results of different methods on the beer dataset. nVAR: number of variables; nLV: number of latent variables; RMSEC: root-mean-square error of calibration set; RMSEP: root-mean-square error of prediction; Q_C^2: coefficient of determination of calibration set; Q_P^2: coefficient of determination of test set; and benchmark results with the form mean value ± standard deviation in 50 runs.

Metrics	PLS	Sparse PLS	lasso	elastic net	fused lasso	OHPL
nLV	5.7±0.7	2.7±1.0	–	–	–	4.5±1.6
nVAR	576	140.4±46.1	26.6±5.8	36.9±9.9	366.2±71.9	76.4±19.5
RMSEC	0.0097±0.0122	0.1370±0.0439	0.1096±0.0280	0.1008±0.0178	0.1067±0.0409	0.1409±0.0283
RMSEP	0.5379±0.1972	0.2092±0.0521	0.2590±0.0771	0.2634±0.0877	0.2417±0.0762	0.1692±0.0338
Q_C^2	0.9999±0.0000	0.9963±0.0021	0.9979±0.0010	0.9982±0.0006	0.9978±0.0013	0.9965±0.0015
Q_P^2	0.9145±0.1889	0.9892±0.0088	0.9809±0.0221	0.9790±0.0302	0.9833±0.0191	0.9929±0.0052

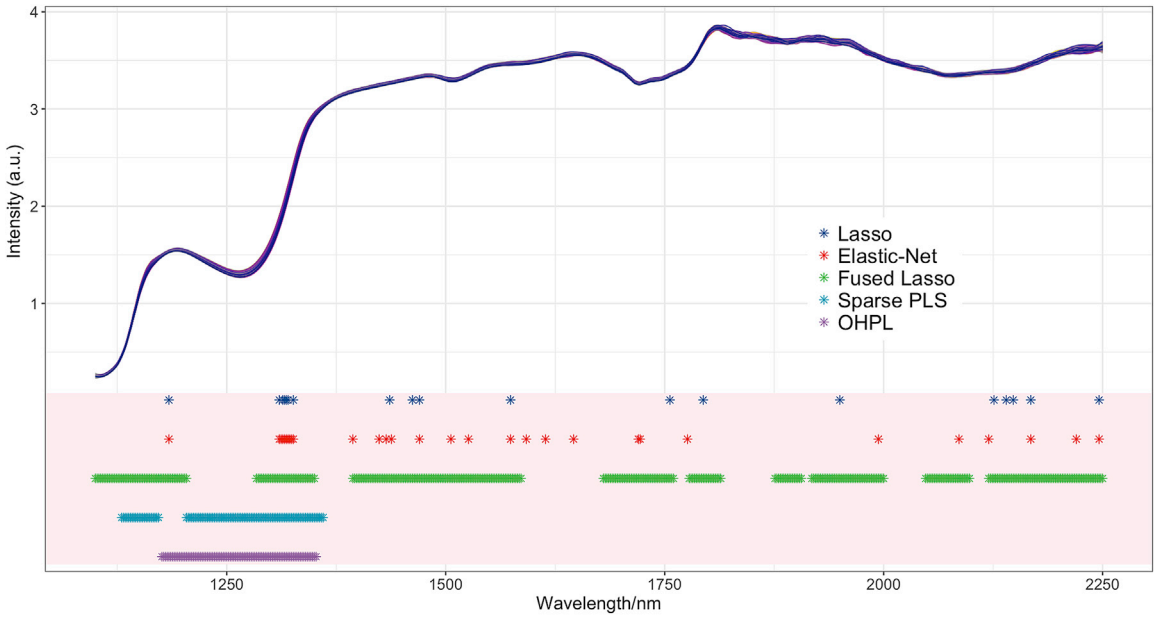

Fig. 2. The original NIR spectra of the beer extract concentration data (top), and a comparison of the variables selected by lasso, elastic net, fused lasso, sparse PLS, and OHPL on the beer dataset (bottom).

which were scanned with an NIR spectrophotometer, and fluorescence excitation-emission matrices (EEMs) were recorded with a spectrofluorometer. In this study, we choose 1100–2500 nm range of NIR (700 data points) as the design matrix **X** based on the original paper [51]. Soil organic matter (SOM) is considered as the property of interest **y**. More details about the data are described in Ref. [51]. Besides, we randomly split the data set into a training data set containing 70% of the samples and a test data set with the remaining samples.

3.4. Software implementation

All the code and experiments are written and implemented in R (https://www.r-project.org). There are several packages for PLS modeling in R, such as the pls package [52] and the enpls package [53]. In this study, PLS is implemented with the pls package [52]. The glmnet package [54] is used to fit lasso and elastic net models. The genlasso package [55] is applied to fit fused lasso models. The spls package [56] provides functions for fitting sparse PLS regression models.

4. Results and discussion

To evaluate the performance of OHPL, some state-of-the-art methods, such as Sparse PLS, lasso, elastic net and fused lasso, are tested on three datasets. The common trait of these methods is using L_1-penalty term to select the important variables. Many high-performance variable selection methods have been developed in the chemometrics community, including but not limited to the following methods: MW-PLS [33,34], CARS [31] and iVISSA [36]. This paper will focus on the discussion of the penalized regression techniques; comparisons of those methods are not considered in this paper. In order to ensure reproducibility and stability of the experiment results, each dataset was randomly split for 50 times to produce different training sets and test sets. Each method was applied to the 50 different training sets, and predicted on the corresponding 50 test sets. The final evaluation result was the average of the results from the 50 runs.

4.1. Beer dataset

Table 1 presents the evaluation results obtained by full-spectrum PLS, Sparse PLS, lasso, elastic net, fused lasso, and OHPL. From Table 1, we can see that the five methods with variable selection effect all have better prediction performance than the full-spectrum PLS on this dataset. Besides, compared to the results of PLS, the RMSEP values of OHPL decreased remarkably from 0.5379 to 0.1692. Moreover, the OHPL method outperforms the other variable selection methods. OHPL has the lowest RMSEP (0.1692), followed by Sparse PLS (0.2092), fused lasso (0.2417), lasso (0.2590) and elastic net (0.2634). OHPL has the highest Q_P^2 (0.9929), followed by Sparse PLS (0.9892), fused lasso (0.9833), lasso (0.9809) and elastic net (0.9790). As described above, OHPL can significantly improve the accuracy of the original lasso, and prediction ability and interpretability of the PLS method.

Fig. 2 Shows the variables selected by different methods on the beer dataset within the 50 experiments. Fused lasso (366.2) and Sparse PLS (140.4) tend to select a larger number of variables, while the nVAR obtained by OHPL (76.4), elastic net (36.9), and lasso (26.6) are lower.

Table 2
The benchmark results of different methods on the wheat dataset. nVAR: number of variables; nLV: number of latent variables; RMSEC: root-mean-square error of calibration set; RMSEP: root-mean-square error of prediction; Q_C^2: coefficient of determination of calibration set; Q_P^2: coefficient of determination of test set; and benchmark results with the form mean value ± standard deviation in 50 runs.

Metrics	PLS	Sparse PLS	lasso	elastic net	fused lasso	OHPL
nLV	9.9±0.3	7.9±0.3	–	–	–	7.6±0.7
nVAR	700	612.1±123.5	12.6±2.4	83.4±53.9	130.5±78.8	84.5±45.1
RMSEC	0.3296±0.0287	0.4006±0.0383	0.6866±0.0560	0.6876±0.0569	0.6678±0.1146	0.2826±0.0615
RMSEP	0.5301±0.0712	0.5819±0.1004	0.7472±0.1381	0.7467±0.1371	0.7344±0.1818	0.2889±0.0310
Q_C^2	0.9100±0.0142	0.8668±0.0290	0.6061±0.0651	0.6049±0.0657	0.6244±0.1070	0.9321±0.0263
Q_P^2	0.7411±0.0838	0.6809±0.1121	0.4976±0.1571	0.4992±0.1540	0.4889±0.2333	0.9115±0.0232

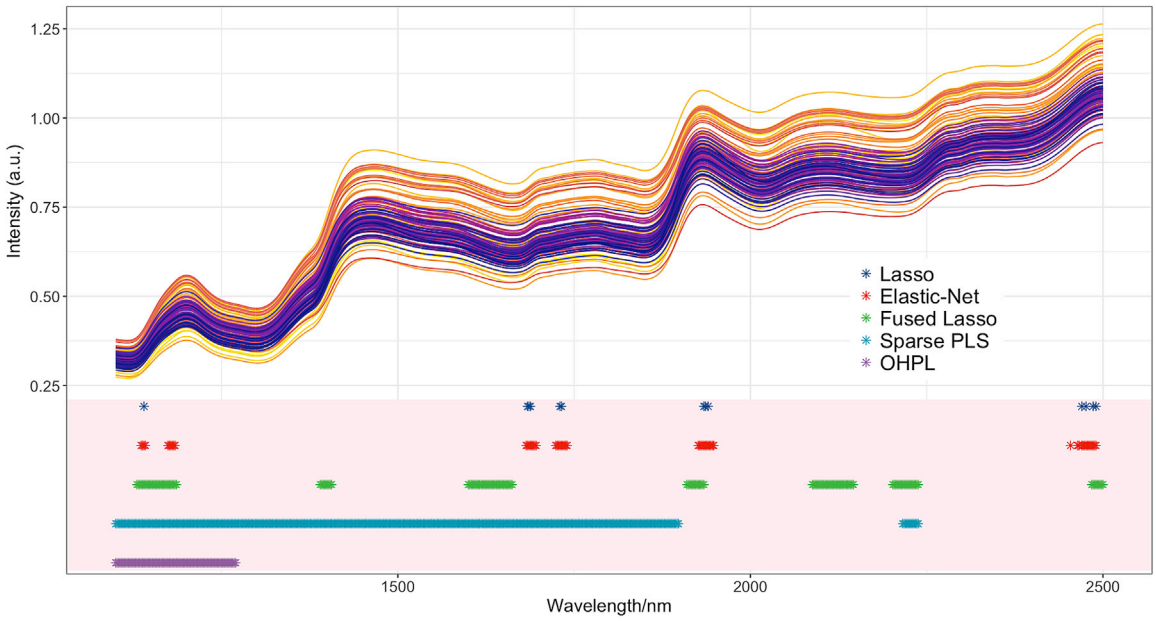

Fig. 3. The original NIR spectra of the wheat protein data (top), and a comparison of the variables selected by lasso, elastic net, fused Lasso, sparse PLS, and OHPL on the wheat dataset (bottom).

From Fig. 2, we can observe that the wavelengths selected by elastic net and lasso are quite similar, whereas the elastic net selected more variables than the lasso. Nine groups selected by fused lasso are located at the region of 1100–1200 nm, 1300–1350 nm, 1400–1590 nm, 1680–1760 nm, 1780–1820 nm, 1878–1908 nm 1920–2000 nm, 2050–2100 nm, and 2122–2250 nm. The intervals selected by Sparse PLS and OHPL are quite similar. The groups selected by Sparse PLS are located at 1132–1174 nm and 1206–1362 nm. The group selected by OHPL is the region of 1172–1352 nm, which corresponds to the first overtone of O—H stretching bond vibration [55].

4.2. Wheat dataset

For the wheat protein dataset, Table 2 lists the results obtained from the different methods. From the table, we can easily observe that the best predictions concerning RMSEP are all achieved by OHPL. Compare to the results of the full spectra, the RMSEC and RMSEP values for OHPL decreased remarkably by 14.3% and 45.5%, respectively. Compare to the results of the lasso, the RMSEC and RMSEP values for OHPL decreases remarkably by 58.8% and 61.3%, respectively. The prediction performance of the original lasso can be significantly improved. OHPL also outperforms other methods significantly with the smallest RMSEC and RMSEP. In addition, OHPL generates the smallest standard deviation for RMSEP, indicating the highest stability on this dataset. Moreover, OHPL shows the highest Q_C^2 (0.9321) and Q_P^2 (0.9115), follows by PLS (0.9100 and 0.7411), Sparse PLS (0.8668 and 0.6809), elastic net (0.6049 and 0.4992), and lasso (0.6061 and 0.4976). The prediction error of fused lasso is the highest with the lowest Q_C^2 and Q_P^2, as displayed in Fig. 5 (B). This example also demonstrates that the generalizations of lasso and OHPL can obtain good performance when dealing with spectral data.

Fig. 3 displays the variables selected by the five variable selection methods on the wheat dataset. SPLS selects the largest number of

Table 3
The benchmark results of different methods on the soil dataset. nVAR: number of variables; nLV: number of latent variables; RMSEC: root-mean-square error of calibration set; RMSEP: root-mean-square error of prediction; Q_C^2: coefficient of determination of calibration set; Q_P^2: coefficient of determination of test set; and benchmark results with the form mean value ± standard deviation in 50 runs.

Metrics	PLS	Sparse PLS	lasso	elastic net	fused lasso	OHPL
nLV	10±0.0	10±0.0	–	–	–	9.9±0.2
nVAR	700	546.2±162.7	24.8±7.8	78.3±42.8	251.6±49.1	392.0±106.4
RMSEC	1.6276±0.1179	1.5955±0.1426	4.030±0.1591	4.0570±0.1413	2.9819±0.3354	1.4766±0.1784
RMSEP	2.2516±0.3543	2.1437±0.3123	4.1248±0.4614	4.1425±0.4735	3.2087±0.3986	1.6533±0.3535
Q_C^2	0.9769±0.0036	0.9771±0.0004	0.8612±0.0158	0.8594±0.0154	0.9241±0.0133	0.9812±0.0004
Q_P^2	0.9464±0.0235	0.9519±0.0036	0.8300±0.0528	0.8289±0.0520	0.8932±0.0473	0.9736±0.0080

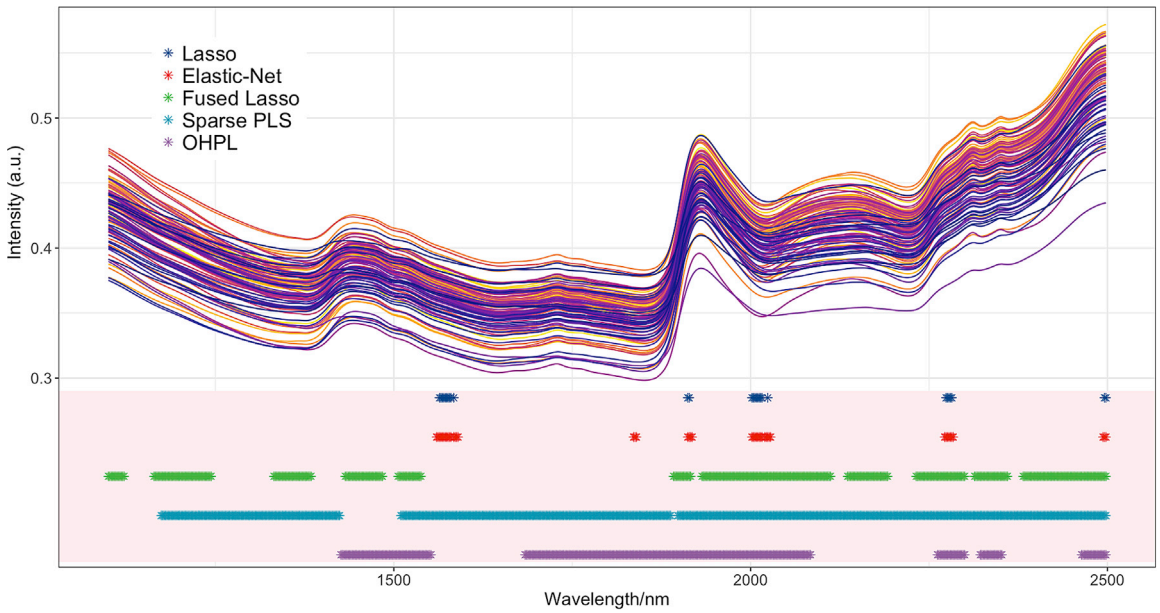

Fig. 4. The original NIR spectra of the soil organic matter data (top), and a comparison of the variables selected by lasso, elastic net, fused Lasso, sparse PLS, and OHPL on the soil dataset (bottom).

variables (612.1), while lasso selects the smallest number of variables (12.6). As mentioned before, the wavelengths selected by elastic net are similar to those selected by lasso, but the wavelengths selected by elastic net are more grouped. The reason is that elastic net takes the L_2-penalty into account, thus encourages the grouping effect. Fused lasso identifies seven intervals in this dataset; they are located in 1132–1188 nm, 1392–1408 nm, 1602–1664 nm, 1912–1938 nm, 2090–2148 nm, and so on. The region 1100–1900 nm and the region 2218–2240 nm are identified by Sparse PLS. The variables selected by OHPL are located in the region 1100–1300 nm, which is consistent with the wavelengths selected by the GA-PLS-RC [57]. The selected informative intervals are distributed in a broad region, which, is implicitly in accordance with the complex structure characteristics of protein. This wide region contains the third overtones of C—H groups (850–865 nm), the second overtones of C—H groups (near 888 nm), the second overtones of O—H groups (972–988), the second overtones of N—H groups (near 1012 nm), and the interactions between them [49].

4.3. Soil dataset

The results on the soil dataset obtained by different methods are illustrated in Table 3 and Fig. 4. It is obvious that the best prediction regarding RMSEP is obtained by OHPL. Moreover, from Table 3, we can observe that elastic net and lasso obtained similar prediction accuracy and there is a clear ranking of prediction accuracy for four alternative methods on this dataset: OHPL > Sparse PLS > PLS > fused lasso. By comparison, lasso has the largest standard deviation of Q_P^2 on this dataset. However, we can find an interesting phenomenon that lasso has a smaller standard deviation of RMSEP than elastic net. Furthermore, another interesting observation is that the standard deviations of RMSEP between PLS and OHPL differ little, while the standard deviations of Q_P^2 differ greatly. SPLS has the smallest standard deviation among five variable selection methods, which means SPLS obtains the highest stability of all. In addition, OHPL has a much smaller standard deviation than lasso. As a summary, Fig. 5 (C) clearly illustrates the fact that the OHPL method is efficient in overcoming the drawbacks of the lasso while improving the prediction accuracy of the original lasso.

Fig. 4 shows the variables selected by different methods on the soil dataset. For this dataset, Sparse PLS and OHPL tend to select a larger number of variables, while the nVAR values obtained by lasso and elastic net are lower. Based on previous studies [50], we know that the informative spectra regions are near 1420 nm (region 1), 1900–1950 nm (region 2), 2040–2260 nm (region 3), and 2440–2460 nm (region 4). These intervals are well in agreement with the major absorption features assigned for organic matter, such as "the absorption at around 1420 nm in the first derivate of the spectra can be attributed to O—H groups in water or cellulose, or to C—H_2 groups in lignin. The absorption at 1900–1950 nm may indicate O—H groups in water or various functional groups present in cellulose, lignin, glucan, starch, pectin and humic acid" [50]. The wavelengths selected by elastic net are similar to those selected by lasso, but the wavelengths selected by elastic net are more concentrated than those selected by lasso. These wavelengths at near 1566 nm, 1914 nm, 2040 nm, 2260 nm, and 2498 nm. Lasso and elastic net failed to select the informative wavelength region 1 and region 3. Fig. 4 shows that fused lasso can select all the informative wavelength region, but it also picks some other wavelength bands, such as 1166–1246 nm and 1334–1386 nm. Sparse PLS selects the largest number of variables. These variables are distributed in three wide bands, such as 1176–1430 nm, 1512–1890 nm, and 1900–2500 nm. In contrast, OHPL selects a smaller number of variables than Sparse PLS. They are near 1420–1554 nm, 1686–2086 nm, 2264–2302 nm, 2324–2354 nm, and 2460–2500 nm. All the selected regions correspond to the chemical bond except the region of 2324–2354 nm. It is noteworthy that the variables of 2324–2354 nm were successfully selected by fused lasso, Sparse PLS, and OHPL, but were failed to obtain by lasso and elastic net. Interestingly, the first three methods (Fused lasso, Sparse PLS, OHPL) have better predictive performance than the last two methods, so it is possible that the region of 2324–2354 nm contains quantitative information for organic matter.

5. Conclusions

In this paper, we proposed a new method named ordered homogeneity pursuit lasso (OHPL) for group variable selection with application to high-dimensional spectroscopy data. OHPL takes the homogeneity structure in high-dimensional data into account. It can be viewed as an improved version of lasso which also has the grouping effect to

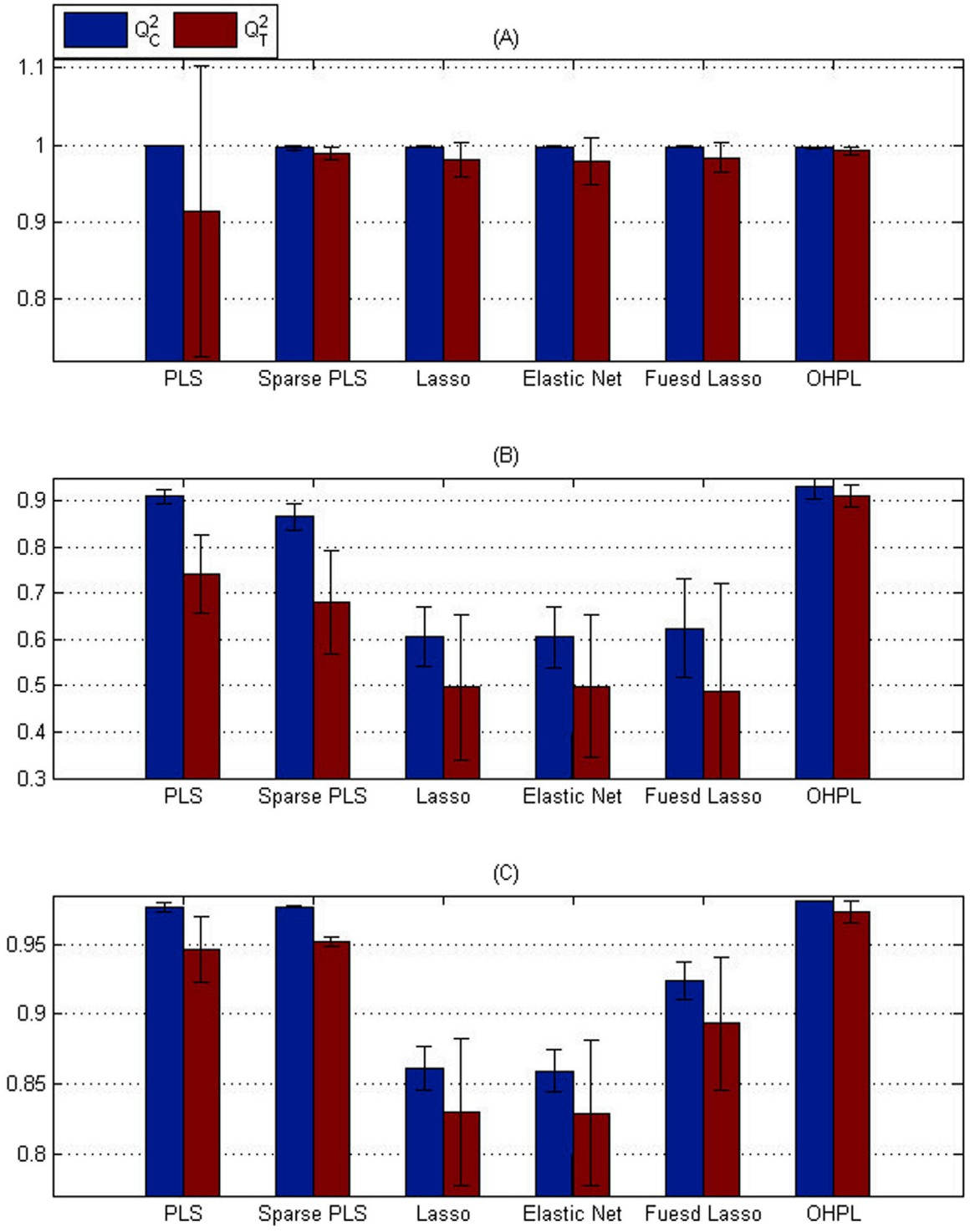

Fig. 5. Q_C^2 and Q_P^2 of PLS, lasso, elastic net, fused lasso, sparse PLS, and OHPL on the three datasets.

automatically select strongly correlated important variables. Some generalizations of lasso, such as elastic net and fused lasso, were proposed to deal with the problems of the original lasso. However, when being applied to three real-world spectroscopic datasets, OHPL outperforms all of them in terms of both predictive performance and accuracy of group variable selection.

Empirical studies on three real-world datasets using different performance metrics showed that OHPL selects more important group variables and has better predictive performance than the other state-of-the-art methods, including Sparse PLS, lasso, elastic net, and fused lasso. Therefore, we believe that OHPL is a promising method for regression in the high-dimensional settings. Although OHPL was only applied to datasets with a natural order in this work, we should point out that it can be certainly used to analyze high-dimensional data with more general group structures. These are beyond the scope of this paper, and our future work will focus on them.

Acknowledgements

We thank the editor and two referees for constructive suggestions that substantially improved this work. This work is financially supported by the National Natural Science Foundation of China (Grant No. 11271374 and 11561010), 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. The Fisher optimal partitions algorithm

To construct the groups of variables, we applied the Fisher optimal partitions (FOP) procedure to the regression coefficients of PLS. This procedure was introduced by W. D. Fisher in 1958 [46]. Based on Hartigan's [58] work, we describe the details of FOP as follows.

Given n ordered objects $a_1, a_2, \cdots, a_n$, we want to split them into K groups, and these groups are constrained to consist of intervals of objects $(a_i, a_{i+1}, \cdots, a_j)$.

Step 1. For the k-th group $G_k = (a_i, a_{i+1}, \cdots, a_j)$, for all i, j, $1 \le i < j \le n$, the diameter $D(i,j)$ is determined by $D(i,j) = \sum_{t=i}^{j}(a_t - \bar{a}_k)^2$, where $\bar{a}_k$ is the mean of the corresponding group.

Step 2. For the number of groups 2, the errors of "optimal partitions" is calculated by

$$L[P(i,2)] = \min_{2 \le h \le i}[D(1,h-1) + D(h,i)] \tag{A-1}$$

where $2 \le i \le n$ and for $i \le n$ and $g \le K$, $P(i,g)$ denotes the optimum g partition of objects $a_1, a_2, \cdots, a_i$.

Step 3. For each g, the errors of the optimal partitions $L[P(i,g)]$ is computed by

$$L[P(i,g)] = \min_{g \le h \le i}[L[P(h-1,g-1)] + D(h,i)] \tag{A-2}$$

where $g \le i \le n$ and $3 \le g \le K$.

Step 4. The optimal partition $P(n,K)$ is searched from the table of errors $L[P(i,g)](1 \le g \le K, 1 \le i \le n)$ by the smallest j satisfying $L[P(n,K)] = L[P(j-1,K-1)] + D(j,n)$. Then the last group is $(j, j+1, \cdots, n)$. Next, find j' such that

$$L[P(j-1,K)] = L[P(j-1,K-1)] + D(j', j-1). \tag{A-3}$$

and the second-to-last group of $P(n,K)$ is $(j', j'+1, \cdots, j-1)$, and so forth.

References

[1] R. Bellman, Adaptive Control Processes, Princeton University Press, Princeton, 1961.
[2] T.T. Cai, X.T. Shen, High-dimensional Data Analysis, Higher Education Press, Beijing, 2010.
[3] H. Martens, T. Næs, Multivariate Calibration, Wiley, New York, 1989.
[4] P. Filzmoser, B. Walczak, What can go wrong at the data normalization step for identification of biomarkers? J. Chromat. A 1362 (2014) 194.
[5] F. Marini, D. de Beer, E. Joubert, B. Walczak, Analysis of variance of designed chromatographic data sets: the analysis of variance-target projection approach, J. Chromat. A 1405 (2015) 94.
[6] M. Daszykowski, J. Orzel, M.S. Wrobel, H. Czarnik-Matusewicz, B. Walczak, Improvement of classification using robust soft classification rules for near-infrared reflectance spectral data, Chemom. Intell. Lab. 109 (2011) 86.
[7] W. Wu, M. Daszykowski, B. Walczak, B.C. Sweatman, S.C. Connor, J.N. Haselden, D.J. Crowther, R.W. Gill, M.W. Lutz, Peak alignment of urine NMR spectra using fuzzy warping, J. Chem. Inf. Model. 46 (2006) 863.
[8] X.X. Zhang, T.D. Johnson, R.J.A. Little, Y. Cao, Quantitative magnetic resonance image analysis via the EM algorithm with stochastic variation, Ann. Appl. Stat. 2 (2008) 736.
[9] H. Chun, S. Keles, Sparse partial least squares regression for simultaneous dimension reduction and variable selection, J. R. Stat. Soc. B 71 (2010) 3.
[10] J.Q. Fan, R.Z. Li, Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery, 2006. Madrid.
[11] R. Tibshirani, Regression shrinkage and selection via the LASSO, J. R. Stat. Soc. B 55 (1996) 267.
[12] Q.S. Xu, Y.Z. Liang, Y.P. Du, Monte Carlo cross-validation for selecting a model and estimating the prediction error in multivariate calibration, J. Chemom. 18 (2004) 112.
[13] M.S. Oh, E.S. Park, B.S. So, Bayesian variable selection in binary quantile regression, Stat. Probab. Lett. 118 (2016) 177.
[14] S.D. Jong, SIMPLS: an alternative approach to partial least squares regression, Chemom. Intell. Lab. 18 (1993) 251.
[15] S. Wold, M. Sjöström, L. Eriksson, PLS-regression: a basic tool of chemometrics, Chemom. Intell. Lab. 58 (2001) 109.
[16] A.E. Hoerl, R. Kennard, Ridge regression: biased estimation for nonorthogonal problems, Technometrics 12 (1970) 55.
[17] I.E. Frank, J.H. Friedman, A statistical view of some chemometrics regression tools, Technometrics 35 (1993) 109.
[18] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, J. R. Stat. Soc. B 67 (2005) 301.
[19] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused lasso, J. R. Stat. Soc. B 67 (2005) 91.
[20] M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables, J. R. Stat. Soc. B 68 (2006) 49.
[21] P. Filzmoser, M. Gschwandtner, V. Todorov, Review of sparse methods in regression and classification with application to chemometrics, J. Chemom. 26 (2012) 42.
[22] J.H. Kalivas, Overview of two-norm (L2) and one-norm (L1) Tikhonov regularization variants for full wavelength or sparse spectral multivariate calibration models or maintenance, J. Chemom. 26 (2012) 218.
[23] P. Shahbazikhah, J.H. Kalivas, E. Andries, T. O'Loughlind, Using the L1 norm to select basis set vectors for multivariate calibration and calibration updating, J. Chemom. 30 (2016) 109.
[24] T. Randolph, J.M. Ding, M.G. Kundu, J. Harezlak, Adaptive penalties for generalized Tikhonov regularization in statistical regression models with application to spectroscopy data, J. Chemom. (2016), http://dx.doi.org/10.1002/cem.2850.
[25] H. Higashi, G.M. ElMasry, S. Nakauchi, Sparse regression for selecting fluorescence wavelengths for accurate prediction of food properties, Chemom. Intell. Lab. 154 (2016) 29.
[26] 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. 150 (2016) 58.
[27] T. Mehmood, B. Ahmed, The diversity in the applications of partial least squares: an overview, J. Chemom. 30 (2016) 4.
[28] D.J. Chung, S. Keles, Sparse partial least squares classification for high dimensional data, Stat. Appl. Gene. Mole. Biol. 9 (2010). Article 17.

[29] W.S. Cai, Y.K. Li, X.G. Shao, A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra, Chemom. Intell. Lab. Syst. 90 (2008) 188.
[30] Q.J. Han, H.L. Wu, C.B. Cai, L. Xu, R.Q. Yu, An ensemble of Monte Carlo uninformative variable elimination for wavelength selection, Anal. Chim. Acta 612 (2008) 12.
[31] H.D. Li, Y.Z. Liang, Q.S. Xu, D.S. Cao, Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration, Anal. Chim. Acta 648 (2009) 77.
[32] Y.H. Yun, W.T. Wang, B.C. Deng, G.B. Lai, X.B. Liu, D.B. Ren, Y.Z. Liang, W. Fan, Q.S. Xu, Using variable combination population analysis for variable selection in multivariate calibration, Anal. Chim. Acta 862 (2015) 14.
[33] 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.
[34] 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.
[35] Y.H. Yun, H.D. Li, L.R.E. Wood, 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, Spectro Acta Part A 111 (2013) 31.
[36] 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.
[37] 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.
[38] Z.T. Ke, J.Q. Fan, Y.C. Wu, Homogeneity pursuit, J. Am. Stat. Assoc. 110 (2015) 175.
[39] X.T. Shen, H.C. Huang, Grouping pursuit through a regularization solution surface, J. Am. Stat. Assoc. 105 (2010) 727.
[40] Y. Ke, J.L. Li, W.Y. Zhang, Structure identification in panel data analysis, Ann. Stat. 44 (2016) 1193.
[41] P. Bühlmann, P. Rütimann, S.V.D. Geer, C.H. Zhang, Correlated variables in regression: clustering and sparse estimation, J. Stat. Plan. Inf. 143 (2013) 1835.
[42] S. Reid, R. Tibshirani, Sparse regression and marginal testing using cluster prototypes, Biostatistics 17 (2016) 364.
[43] N. Xiao, Q.S. Xu, Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection, J. Stat. Comput. Simul. 85 (2015) 3755.
[44] 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.
[45] F. Wang, L. Wang, P.X.K. Song, Fused lasso with the adaptation of parameter ordering in combining multiple studies with repeated measurements, Biometrics 72 (2016) 1184.
[46] W.D. Fisher, On grouping for maximum homogeneity, J. Am. Stat. Assoc. 53 (1958) 789.
[47] J.Q. Fan, J.C. Lv, Sure independence screening for ultrahigh dimensional feature space (with discussion), J. R. Stat. Soc. B 70 (2008) 849.
[48] N. Chirico, P. Gramatica, Real external predictivity of QSAR models: how to evaluate it? Comparison of different validation criteria and proposal of using the concordance correlation coefficient, J. Chem. Inf. Model. 51 (2011) 2320.
[49] L. Norgaard, A. Saudland, J. Wagne, 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.
[50] J. Kalivas, Two data sets of near infrared spectra, Chemom. Intell. Lab. 37 (1997) 255.
[51] 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.
[52] B.H. Mevik, R. Wehrens, The pls package: principal component and partial least squares regression in R, J. Stat. Soft. 18 (2007) 1.
[53] N. Xiao, D.S. Cao, M.Z. Li, Q.S. Xu, Enpls: Ensemble Partial Least Squares Regression, R package version 5.7, 2017. URL, https://CRAN.R-project.org/package=enpls.
[54] J. Friedman, T. Hastie, R. Tibshirani, Regularization paths for generalized linear models via coordinate descent, J. Stat. Soft. 33 (2010) 1.
[55] T.B. Arnold, R.J. Tibshirani, Efficient implementations of the generalized lasso dual path algorithm, J. Comput. Graph. Stat. 25 (2016) 1.
[56] D.J. Chung, H. Chun, S. Keles, Spls: Sparse Partial Least Squares (SPLS) Regression and Classification, R package version 2.2-1, 2013. URL, https://CRAN.R-project.org/package=spls.
[57] D. Li, Z. Wu, D. Xu, Y. Xu, Measurement of the principal components in beer by means of near infrared spectroscopy, Chin. J. Anal. Chem. 32 (2004) 1070.
[58] J.A. Hartigan, Clustering Algorithms, John Wiley & Sons, New York, 1975.