
Title	A clustering group lasso method for quantification of adulteration in black cumin seed oil using Fourier transform infrared spectroscopy
Author(s)	Ying Zhu, Lin Zou, and Tuck Lee Tan

Copyright © 2022 Elsevier

This accepted manuscript is made available under the CC-BY-NC-ND 4.0 license <http://creativecommons.org/licenses/by-nc-nd/4.0/>

The final publication is available at: <https://doi.org/10.1016/j.chemolab.2021.104471>

A Clustering Group Lasso Method for Quantification of Adulteration in Black Cumin Seed Oil Using Fourier Transform Infrared Spectroscopy

Ying Zhu ^{a,*}, Lin Zou^a, Tuck Lee Tan^a

^aNational Institute of Education, Nanyang Technological University
1, Nanyang Walk, Singapore 637616

Email: *ying.zhu@nie.edu.sg

Date: 28th October 2021

*Corresponding Author:

Dr. Ying Zhu

Postal Address:

National Institute of Education

Nanyang Technological University

1, Nanyang Walk

Singapore 637616

Tel: (65) 6790 3989

Fax: (65) 6896 9417

Email: ying.zhu@nie.edu.sg

No. of pages: 28

No. of figures: 10

No. of tables: 2

ABSTRACT

Black cumin seed oil (BCSO) contains a large number of bioactive compounds and thus has many medicinal health benefits and uses. Its high economic profits in the market lead to the frequent occurrence of adulterating this oil with cheaper edible oils such as grape seed oil, walnut oil. It is difficult to detect adulteration as the oil adulterant has similar physical characteristic and even similar chemical composition to the authentic oil. The development of an accurate and rapid analytical method using attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopy is of essential importance for determination of authenticity of BCSO and quantification of oil adulterants. In this study, BCSO and grape seed oil (GSO) were mixed in various ratios to mimic the adulteration. A clustering group lasso method was developed by incorporating both the high correlation structure of spectral variables and the underlying group features into the model. Instead of assuming that groups are known a priori as does ordinary group lasso, the clustering group lasso infers groups of spectral features from the data and encourages spectral variables within a group to have a shared association with the response. The model using ATR-FTIR spectroscopy proved to be a powerful tool to quantify BCSO adulteration with high accuracy and can accurately predict the quantity of adulterant at levels as low as 5%. With a substantial reduction in number of spectral features, the clustering group lasso model shows a simple regression coefficient profile with improved interpretability as compared to the ordinary group lasso model and other penalized models. The spectral regions automatically selected for quantification of BCSO adulteration can be helpful for the interpretation of the major chemical constituents of BCSO regarding its anti-cancer and anti-inflammatory effects from a chemometric perspective.

Keywords: *Black cumin seed oil, oil adulteration, clustering, group lasso, elastic net, variable selection, penalized regression.*

1. Introduction

Black cumin (*Nigella sativa*), which belongs to the botanical family of Ranunculaceae, commonly grows in the Southern Europe, Middle East, and Southwest Asia. For centuries, seeds of black cumin have traditionally been used in the Asian and Middle East countries for the promotion of good health and the treatment of many diseases, such as asthma, bronchitis, rheumatism and other inflammatory diseases. Black cumin seed oil (BCSO), as a natural remedy and dietary supplement, is one of the most extensively studied oils in recent years due to its many beneficial medicinal effects, including antibacterial, antitumor, anti-oxidant and anti-inflammatory [1-3]. As a medicinal oil, the chemical composition of BCSO is very rich and diverse. It contains amino acids, proteins, carbohydrates, fixed and volatile oils [4]. The main bioactive constituent of the volatile oil of the BCSO is Thymoquinone, which is known as a potent anti-inflammatory agent and shows promise in treating epilepsy, allergies, and boosting immune system [1,3].

In the market, BCSO is about 10–15 times more expensive than other edible oils such as grape seed oil, walnut oil and soybean oil. Adulteration of BCSO with other cheaper oils thus occurs frequently to gain economic profits. Oil adulteration is a serious problem which may cause dangerous effects, such as emergence of an allergic reaction. The need for oil authentication is thus a necessity of the food industry. It is difficult to detect adulteration when the oil adulterant has similar physical characteristic and even similar chemical composition to the authentic oil [5,6]. Various chemical methods such as high-performance liquid chromatography (HPLC), carbon isotope ratio, have been developed for detection of oil adulteration. However, those methods generally require the destruction of sample materials, and are too laborious, time-consuming and expensive. Therefore, the development of an accurate, rapid and simple analytical method is of essential importance to detect and quantify the oil adulterants.

The attenuated total reflection Fourier transform infrared (ATR-FTIR) spectroscopic technique allows direct, accurate and rapid measurement of spectra of the materials for authentication study of edible oils. Using FTIR, as there is no need of chemical processing or treatment to the sample materials, the use of hazardous solvents and chemical reagents can be avoided and its application to analysis of edible oils is thus considered as “green analytical technique” [7].

ATR-FTIR spectra of edible oils consist of many broad and overlapping absorption bands associated with vibrational modes featuring different molecular functional groups of the compounds, which are sensitive to the physical and chemical states of the compounds and can be detected at low levels [8]. From literature reviews, grape seed oil (GSO) has most similar fatty acid profiles with BCSO among many plant oils [9,10] and is thus considered as potential oil adulterant in BCSO in this study. Due to the similar chemical composition between authentic oil and its adulterant, only very subtle differences are visible in the ATR-FTIR spectra of BCSO adulterated with GSO at various ratios. It is increasingly difficult to predict the quantity of the BCSO adulteration with proportion of GSO adulterant lower than 20%. Therefore, even for most experienced analysts, it is practically impossible to detect and quantify the adulterants by a visual inspection or routine analytical approaches. Considering the complex nature of FTIR spectra, modern

multivariate statistical methods in chemometrics have been appropriately applied in this study to extract the most informative features contained in the spectroscopic data, allowing for detection and quantification of GSO adulterant in BCSO with subtle compositional differences among samples. Although ATR-FTIR spectral features are highly dependent due to the intrinsic complexity of biological systems, quite often, only a small subset of spectral features may have a significant impact on response variable while majority of the features may be irrelevant or redundant to the regression or discrimination [11,12]. Identifying the most important spectral variables is particularly helpful in interpreting spectral features and understanding the major chemical constituents of BCSO. Furthermore, there exist group structures in ATR-FTIR spectra of edible oils corresponding to their underlying molecular functional groups present in the compounds. Ordinary variable selection methods fit the model based on individual predictor variables, ignoring the group structure of spectroscopic data when extracting the most informative spectral features. The significance of these problems arouses our interest in determining if a small subset of variables selected in groups out of thousands of strongly correlated spectral variables obtained from only hundreds of samples could contain as many important features as the whole spectrum does with the aim of oil authentication and model interpretation.

Recent studies have shown that FTIR spectroscopy, as a rapid and nondestructive authentication tool, is capable of detecting and quantifying adulteration of various edible oils when combined with multivariate analysis approaches. Though the usual practice of identifying adulteration involves both detection and quantification of adulteration, the work presented in this paper, as a preliminary study, focuses on a quantification problem—aiming at predicting percentage of BCSO pureness, given that the adulterant oil type is known. Thus the models explored in this paper are in regression setting. The commonly used multivariate regression methods, principal component regression (PCR) and partial least squares regression (PLSR) [13-15], are able to tackle the problem due to high-dimensional and collinear spectroscopic data by projecting the data into a low-dimensional space spanned by latent variables without significant loss of information. However, the interpretation for those traditional dimension reduction methods is often challenging. Sparse methods like sparse PCR [16] and sparse PLSR [17,18], have been proposed to incorporate sparsity into the projection directions, thus performing feature selection and dimension reduction simultaneously to improve model interpretability. However, its L_1 -norm penalty penalizes each latent variable (component) independently and thus different sets of predictor variables may be selected for different components. This may lead to a relatively large number of variables selected for the model, as compared to those methods which provide a global selection of variables. Similar to PCR and PLSR, the constructed components represent combinations of the original spectral variables, and thus may not be straightforward to interpret. In recent years, various penalized methods in regression context have been developed for high-dimensional and collinear data in order to implement effective variable selection in an attempt to reflect grouping effects. These methods provide a global selection of variables instead of doing selection for each latent variable. The elastic net proposed by Zou and Hastie [19] with a combination of both L_1 and L_2 -norms penalties, achieves model parsimony as well as encourages grouping effects. However, the grouping effects in elastic net are not obvious due to the fact that its penalty does not explicitly

accommodate correlation structure of predictor variables. The group lasso method proposed by Yuan and Lin [20] can solve this problem by applying penalty to groups of predictor variables given that the distinct groups are known through prior or external information about the predictors. The problem is that if such grouping information is not available and it is often the case for spectroscopic data, we may wish to identify the groups from the data. It would be more informative to exploit these groups information when performing penalized regression or discrimination of samples. Little advances have been made on their application to chemometrics tasks, and particularly there are very few examples on development of group lasso method on spectroscopic data for food authentication.

In this paper, a clustering group lasso method was thus developed by encouraging correlated spectral features within a group to have a shared association with the response for quantification of adulteration in BCSO. Instead of assuming that groups are known a priori as does ordinary group lasso, the clustering group lasso infers clusters (groups) of features from the data and then incorporates the spectral correlation structures into the regression model for shrinking and selecting the predictor variables. Four penalized regression models including sparse PLSR, elastic net, ordinary group lasso and clustering group lasso were investigated to identify important spectral features for quantification of GSO adulterant in BCSO.

2. Materials and Methods

2.1 Sample preparation

Genuine BCSO and GSO were purchased from a Singapore health product store and a local supermarket respectively. In order to simulate the adulteration of BCSO, the BCSO and GSO were mixed by pipetting the oils to a tube with BCSO at different percentages of 10, 30, 50, 60, 70, 90 and 95 as shown in Table 1. For each mixing ratio the two types of oil were mixed three times, and for each mixture four drops (samples) were taken from different layers of mixed oil. One bottle of BCSO and one bottle of GSO were used for mixing the oils. In addition, pure oil samples taken from genuine BCSO and GSO were used in our study as shown in Table 1. The ATR-FTIR spectra were measured directly from the samples without further processing. Each oil sample was scanned two or three times. In total 286 spectra were collected consisting of 63 spectra from 21 genuine BCSO samples, 60 spectra from 20 genuine GSO samples and 163 spectra from 82 mixed oil samples. These spectra were used as training set to derive regression rules which will be described later in the Statistical Analysis Section.

An independent spectral data set was collected from another 48 samples of oil and was not used until after the regression models had been trained. This data set included 10 samples from genuine BCSO, 10 samples from genuine GSO and 28 samples from mixed oil. The oil samples were mixed independently of the training set by using another two bottles of BCSO and GSO. For each mixing ratio the two types of oil were mixed once and for each mixture four drops (samples) were taken from different layers of the mixed oil.

With each sample measured two or three times in the same way as in the training set, in total 116 spectra including 30 genuine BCSO spectra, 30 genuine GSO spectra and 56 mixed oil spectra were collected as a test set for the use of prediction in Section 2.4

Table 1. Mixtures of BCSO with GSO in various mixing ratios and corresponding number of samples and number of spectra for training set.

Mixing ratio (BCSO: GSO)	Number of samples	Number of Spectra
0:100	20	60
10:90	12	24
30:70	12	24
50:50	12	23
60:40	11	22
70:30	12	24
90:10	12	24
95:5	11	22
100:0	21	63
Total	123	286

2.2 Spectral acquisition

A Perkin-Elmer Spectrum 100 model FTIR spectrometer equipped with an attenuated total reflection (ATR) accessory was used to obtain the ATR-FTIR spectra for our study. The oil sample was placed, using a Pasteur pipette, directly in contact with the ATR ZnSe crystal at a controlled ambient temperature (25°C). For all the genuine oil samples and the mixed oil samples, each ATR-FTIR spectrum was measured by 30 scans over the mid-IR region of 4000-550 cm^{-1} at intervals of 1 cm^{-1} . For each sample spectrum, a signal-to-noise ratio higher than 40 is obtained by an average of the 30 scans. A background spectrum through the clean ZnSe crystal was first acquired before placing the sample onto the surface of the crystal. Each sample spectrum was then automatically ratioed against the background spectrum to produce an absorbance spectrum for data analysis. Using ratioed sample spectra eliminated the interference from the absorptions due to atmospheric water vapor and carbon dioxide [21]. At the end of spectral measurement for each sample, the ATR plate surface was carefully cleaned and dried with a soft lab tissue before filling in with the next sample.

2.3. Spectral pre-treatment

ATR-FTIR spectral pattern is significantly influenced by both chemical compositions and physical properties of the sample. The variations related to physical effects account for majority of the variations among the spectrum, while the variations associated with chemical composition are considered to be small and may be masked by the physical variations. Conducting appropriate mathematical pre-treatments is essential for reducing variations due to physical effects for the purpose of enhancing the contribution of the chemical composition [22].

The spectral pre-treatment used in this study include a Savitzky-Golay smoothing [23] step with a polynomial of degree 4 and a window size of 11 points, a spectral points reduction step with every third point selected to improve computation speed, a cropping step with regions of low signal-to-noise ratios removed, followed by a standard normal variate (SNV) normalization step [24] where every spectrum was standardized to have a mean of zero and a unit standard deviation in order to correct light scatter due to different particle sizes. After pretreatment, only the spectral region between 3999 and 556 cm^{-1} , with 1149 wavenumber points, were used to analyze data. In Figure 1, the mean FTIR spectra of genuine BCSO, genuine GSO, and BCSO in a mixture with GSO in various mixing ratios all show very similar patterns. If one examines the spectra closely, some subtle differences are revealed among the BCSO spectrum, the GSO spectrum and the mixed oil spectra which can be observed in the regions of (a) 3000-2800 cm^{-1} (b) 1800-1600 cm^{-1} (c) 1200-1000 cm^{-1} , and (d) 800-600 cm^{-1} as illustrated in Figure 2. In Figure 2(b) the magnified spectra of mixed oil with high percentage of BCSO have two peaks in the region of 1750-1710 cm^{-1} while the mixed spectra with low percentage of BCSO reveal one peak.

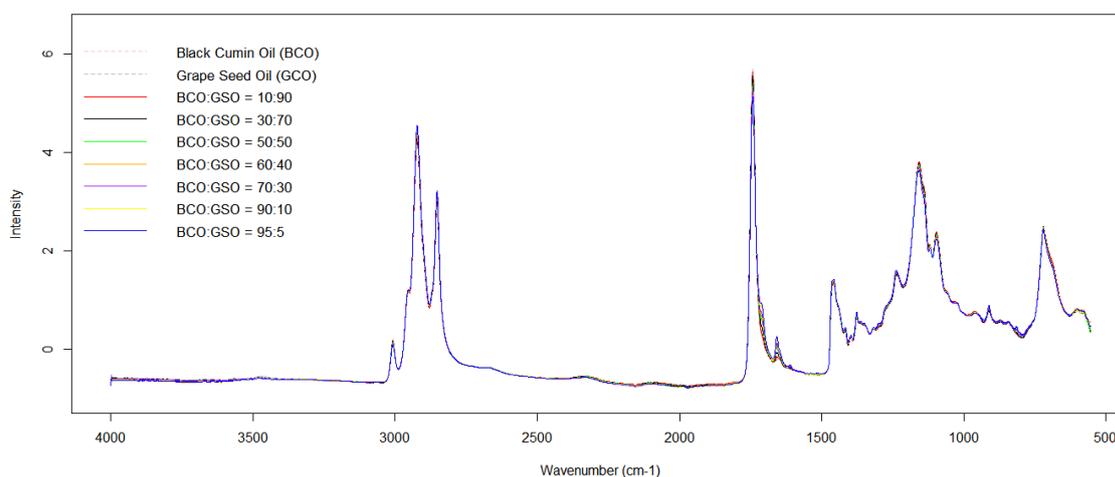


Figure 1: Mean spectra from genuine BCSO, genuine GSO, and BCSO in a mixture with GSO in various mixing ratios at wavenumbers 3999-556 cm^{-1} after pre-treatment.

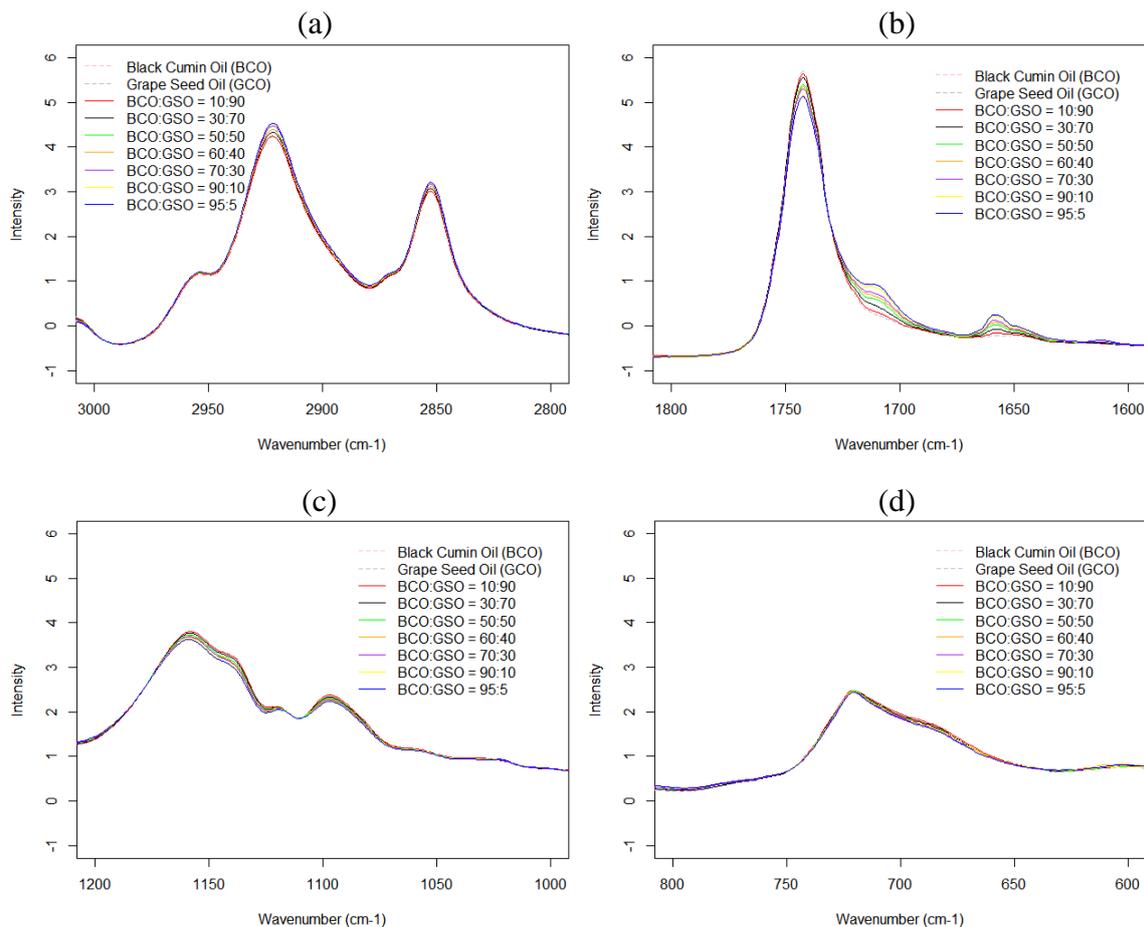


Figure 2. The magnified mean spectra from genuine BCSO, genuine GSO, and BCSO in a mixture with GSO in various mixing ratios in the regions of (a) 3000-2800 cm⁻¹ (b) 1800 -1600 cm⁻¹ (c) 1200-1000 cm⁻¹ (d) 800-600 cm⁻¹.

2.4. Statistical analysis

2.4.1 PLSR and Sparse PLSR

PLSR [14,15] takes into account the relationship between the spectral variables and the response variable for latent variable design. Although PLSR can deal with ill-posed problems and improve the prediction accuracy, latent variables of PLSR have contributions from all the variables and the model interpretation becomes difficult in the presence of large numbers of noise variables.

Motivated by the observation that when noise variables enter the PLSR via direction vectors they would attenuate estimates of the regression parameters, sparse PLSR [18] was proposed for getting sparse solution by imposing penalty on the directions vectors. Let y be a vector of response that denotes the percentage of BCSO in the mixed oil, and X be a matrix of spectral observations. Sparse PLSR finds its first sparse component by solving the following optimization problem:

$$\min_{w,c} \{-\kappa w^T M w + (1 - \kappa)(c - w)^T M (c - w) + \lambda_1 \|c\|_1 + \lambda_2 \|c\|_2^2\}, \quad \text{s.t. } w^T w = 1. \quad (1)$$

where $M = X^T y y^T X$, w and c are the direction vectors, λ_1 is a non-negative tuning parameter. The L_1 -norm penalty is imposed onto a surrogate of the direction vector c instead of the original direction vector w . The two vectors w and c are kept close to each other when calculated by an iteration algorithm. The L_1 -norm encourages sparsity on c whereas the L_2 -norm addresses the potential singularity in M when solving for c . As a large λ_2 value is required to solve problem (1), the λ_2 parameter is set to ∞ to yield the soft threshold estimator which depends only on λ_1 as key tuning parameter [18]. It is noted that this problem becomes the original maximum eigenvalue problem of PLSR when $\kappa = 1$. So we can consider Equ. (1) as a general form for both PLSR and sparse PLSR. Sparse PLSR, as an iteration algorithm, finds first direction vector firstly, then the second and until up to finding k weight vectors for k number of latent components. For each fixed number of components, the optimum parameter λ_1 is chosen as the one that gave the cross-validated mean squared error within one standard error of the minimum.

2.4.2 Elastic Net

Elastic net [19] is a penalized least squares method with a combination of L_1 and L_2 -norms penalties imposed on the regression coefficients. Assume that a spectral data set consist of n spectral observations with p wavenumber points for each spectrum. Let $y = (y_1, \dots, y_n)^T$ be the response variable, where each variable y_i represents the percentage of BCSO in the mixed oil for the i th spectrum. In this study y_i takes the values of 0, 0.1, 0.3, 0.5, 0.6, 0.7, 0.9, 0.95 and 1 for various percentages of BCSO in the mixed oil as shown in Table 1, with the two specific values, 1 and 0, representing the spectra from pure BCSO and pure GSO respectively. Let $x_i = (x_{i1}, \dots, x_{ij}, \dots, x_{ip})^T$, $i = 1, \dots, n$, be the i th observation with p spectral points, where x_{ij} denotes the intensity of the j th spectral point (j th predictor) for the i th spectrum. Let $\beta = (\beta_1, \dots, \beta_p)^T$ be the coefficient vector of p predictors excluding the intercept β_0 . The elastic net solves the following optimization problem:

$$\min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^n (y_i - \beta_0 - x_i^T \beta)^2 + \lambda_2 \sum_{j=1}^p \left[\lambda_1 |\beta_j| + \frac{1}{2} (1 - \lambda_1) \beta_j^2 \right] \right\}, \quad (2)$$

where λ_1 ($0 \leq \lambda_1 \leq 1$) as a tuning parameter controls the distribution between the two penalty terms, and the non-negative tuning parameter λ_2 controls the general degree of penalization. Among the two penalty terms, the L_1 -norm lasso [25] penalty induces sparsity in coefficients β , and the L_2 -norm ridge regression [26] penalty produces coefficients shrinkage of correlated predictor variables towards each other. As a compromise between the ridge regression penalty ($\lambda_1 = 0$) and lasso penalty ($\lambda_1 = 1$), the elastic net penalty shows advantage in providing shrinkage and variable selection simultaneously. For each fixed λ_2 , when λ_1 increases from 0 to 1, the model sparsity (defined as the number of non-zero coefficients) increases gradually from 0 to the lasso sparsity level. Another advantageous property of the elastic net is that the regression coefficients of strongly

corrected variables have a tendency to be equal (in the absolute value sense) so as to encourage a group effect. The two tuning parameters λ_1 and λ_2 were chosen by first picking a sequence of λ_1 values from 0 to 1 by increment of 0.1, followed by choosing the optimal value of λ_2 for each fixed λ_1 using cross-validation on the training data. A sequence of λ_2 values used in the model span a log-scaled range from -8 to 1 by increment of 0.1.

2.4.3 Group Lasso

Though the elastic net model claims that it encourages a grouping effect, it often shows poor performance of choosing correlated variables in group, particularly when correlations among variables in the same group are non-extreme. Moreover, as elastic net penalty was not designed explicitly for correlated variables, it can hardly deal with spectral features with complex correlation structure. The group lasso [20] can handle these issues with extended lasso penalty to capture structure of grouped variables. When the penalty is applied to each group of variables instead of individual variables, group lasso removes a set of grouped predictor variables by shrinking their coefficients to zero and selects a set of important groups of variables that produce accurate prediction.

Let y be a $n \times 1$ vector of response that denotes the percentage of BCSO in the mixed oil, and X a $n \times p$ matrix of spectra with p spectral features (predictors). Suppose that the p predictors are divided into K distinct groups with p_k the number of predictors in group k , $k = 1, \dots, K$. We use a matrix $X^{(k)}$ to represent a submatrix of X whose columns are the predictors in group k , with $\beta^{(k)}$ the corresponding coefficient vector excluding the intercept β_0 . The group lasso solves the following convex optimization problem:

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - \beta_0 \mathbf{1} - \sum_{k=1}^K X^{(k)} \beta^{(k)}\|_2^2 + \lambda \sum_{k=1}^K \sqrt{p_k} \|\beta^{(k)}\|_2 \right\}, \quad (3)$$

where λ is a non-negative tuning parameter, $\mathbf{1}$ is a vector with all components being one, and the $\|\cdot\|_2$ in the penalty term is the non-squared Euclidean norm. The multiplier $\sqrt{p_k}$'s in the penalty term are used to adjust for the group sizes so that smaller groups would not be overwhelmed by larger groups. The non-squared L_2 norm penalty function used in group lasso is intermediate between the L_1 penalty used in lasso and the L_2 penalty used in ridge regression. Using a non-squared L_2 norm regularization, the group lasso performs variable selection like lasso, but induces sparsity at group level. For individual variables within each group, it acts like ridge. When we increase λ , the model sparsity level increases accordingly with reduced number of non-zero spectral variables being selected, and an entire group of predictors may drop out of the model. In a special case when the group sizes are all equal to one, the group lasso reduces to the lasso.

However, group lasso is designed to select pre-defined groups of predictors. More precisely, prior knowledge on groups of correlated variables is needed for group lasso model. In literature, grouping information is often introduced into a model as a priori utilizing scientifically meaningful knowledge [27]. In the case where intrinsic features and

underlying group structure are not available, we may wish to identify the groups from the data, that is, to partition the p predictor variables into K groups before carrying out group-lasso. In this article we considered two approaches for grouping the predictor variables. One approach, for use with ordinary group lasso in this study, partitioned the spectral variables into K disjoint equal-sized groups, with each group composed of adjacent spectral variables (wavenumbers) considering that the adjacent spectral variables are correlated. Different group size of 10, 20, 30, 40 and 50 spectral variables were tried for grouping the predictors before performing the group lasso. For each group size, the optimal value λ was chosen by cross-validation. The other approach based on hierarchical clustering for use in clustering group lasso will be introduced in the next section.

2.4.4 Clustering Group Lasso

We assume that there are K distinct (but unknown) groups of variables, with moderate or high levels of correlation among the variables within each group, and little or no correlation between the groups. As the underlying group structure of spectral features is unknown, the clustering group lasso is implemented in the following two stages.

In stage 1, we perform a hierarchical clustering based on spectral variables correlation [28, 29] to identify the groups. The aim of the clustering is to find a partition of predictor variables $P_K = \{C_1, C_2, \dots, C_K\}$ such that the variables of $X^{(k)}$ within each cluster C_k are strongly related to each other. This is solved by maximizing the homogeneity function as follows:

$$\max_{P_K} \sum_{k=1}^K H(C_k), \quad (4)$$

where $H(C_k) = \sum_{x_j \in C_k} r_{x_j, s_k}^2$ is a homogeneity function of a cluster C_k that measures correlation between the variables in the cluster C_k and its central synthetic variable s_k that is given by

$$s_k = \operatorname{argmax}_{\mathbf{u} \in R^n} \left\{ \sum_{x_j \in C_k} r_{\mathbf{u}, x_j}^2 \right\}, \quad (5)$$

where r^2 denotes the squared Pearson correlation. The homogeneity function $H(C_k)$ is maximized when all the variables in the cluster are strongly correlated to s_k . It can be shown that synthetic variable s_k is the first principal component of $X^{(k)}$ in the cluster C_k [28, 30]. From this point of view, the synthetic variables as latent components may look like sparse PCA components for feature selection and clustering. However, in the case when sparse PCA loadings overlap, it is difficult to cluster the variables [31].

This clustering approach based on homogeneity criterion, though only involving quantitative predictor variables in this study, can be easily extended to allow both quantitative and qualitative variables or a mix of those variables [32].

Detailed algorithm of the hierarchical and partitioning clustering to solve problem in Equ.(4) is given below:

- (1) In step $l = 0$, start with an initial partition in p clusters.

- (2) In step $l = 1, \dots, p - 2$, aggregate two clusters C_i, C_j of the partition in $p - l + 1$ clusters with the smallest dissimilarity function d to obtain a new partition in $p - l$ clusters:

$$d(C_i, C_j) = H(C_i) + H(C_j) - H(C_i \cup C_j).$$

- (3) In step $l = p - 1$, stop the algorithm when the partition in one cluster is achieved.

In stage 2, we hold the partition C_1, C_2, \dots, C_K fixed and β is then estimated by solving

$$\min_{C_1, \dots, C_K, \beta} \left\{ \frac{1}{2n} \left\| y - \beta_0 \mathbf{1} - \sum_{k=1}^K X^{(k)} \beta^{(k)} \right\|_2^2 + \lambda \sum_{k=1}^K \sqrt{p_k} \left\| \beta^{(k)} \right\|_2 \right\} \quad (6)$$

where $\{C_1, \dots, C_K\}$ denotes a partition into K clusters of the p predictor variables, such that $C_j \cap C_l = \emptyset$ if $j \neq l$ and $C_1 \cup C_2 \cup \dots \cup C_K = \{1, \dots, p\}$. Given the partition of predictor variable obtained in stage 1, stage 2 works as the same as group lasso problem in Equ.(3) to estimate β by encouraging variables within a given cluster (group) to take on similar coefficient estimates.

Spectral variables were grouped by using hierarchical clustering with number of clusters K varying from 20 to 90 by increment of 10. For each given number of clusters, the optimal value of λ was chosen by cross-validation.

2.5 Validation and Prediction

For all the models explored in this article, leave-out-one-section cross-validation was used to assess the model performance. Based on the idea of leave-one-out cross-validation, the leave-out-one-section cross-validation trained the algorithm by deriving the regression rules on all the training data except one section of spectral data which was obtained from samples with one particular mixing ratio to be tested.

In this study, from each loop the section left out for testing and the remaining sections for training, though taken from the same bottles, consist of the samples obtained from different mixtures to prevent bias being introduced through non-independence of data. The process was repeated until all the sections with various mixing ratios were tested, and an overall model accuracy (being measured by cross-validated root mean squared error) was thus determined.

The leave-out-one-section cross-validation was used to choose the optimal value of parameter(s) that gave the most regularized model such that the cross-validated mean squared error falls within one standard error of the minimum.

To compare the model robustness, a separate test set described in Section 2.1 was used to make predictions. The same mathematical pre-treatments were first carried out on the test set as described in Section 2.3. Each of the regression models derived from the training set, as mentioned in Section 2.4.1–2.4.3, was then applied to the pre-treated test data for prediction. Though the test set collected involves the same percentages of adulterant in the mixture as the training set does, the explored regression models in this article can be used

to predict the samples with any other different percentages not included in the training set, as long as the samples are within the calibration range.

The R statistical programming language [33] was used to implement all the algorithms for data analyses and computations. The elastic net and group lasso models were executed by using the `glmnet` and `gglasso` packages in R. The hierarchical clustering was implemented by using the `ClustOfVar` package [32]. The PLSR and sparse PLSR models were implemented by using the R packages `plsgenomics` and `spls`.

3. Results and discussion

3.1 Absorption band assignments of ATR-FTIR spectra of BCSO and GSO

Figure 3(a) presents the typical ATR-FTIR spectra from genuine BCSO and GSO after standard pre-treatment was carried out in the region of 3999-556 cm^{-1} . The major infrared (IR) absorption peaks were labelled on the mean spectra of BCSO and GSO. Table 1 lists the spectral wavenumbers with chemical assignment of their corresponding IR absorption bands in BCSO and GSO found in literature [34,35].

BCSO contains both fixed and volatile oil responsible for many health benefits. Fixed oil consists of appreciable amounts of fatty acids (mainly linoleic, oleic and linolenic acids). Thymoquinone, as an important constituent in the volatile oil [36], has been demonstrated to be the major active substances in BCSO [37, 38] and has been found to exhibit anticancer and anti-inflammatory activities [34, 39, 40].

In literature, FTIR spectrum of thymoquinone shows characteristic absorption bands at $\sim 3009 \text{ cm}^{-1}$, $\sim 2924 \text{ cm}^{-1}$ and $\sim 2854 \text{ cm}^{-1}$ that can be assigned to C-H stretching of vinylic group and aliphatic group [40]. Another remarkable absorption band observed at $\sim 1659 \text{ cm}^{-1}$ belongs to the C=O carbonyl stretching of thymoquinone. In this study, though most of these characteristic absorption peaks of thymoquinone appeared in the spectra of both oils, some apparent differences were observed either in the peak intensities especially at $\sim 2924 \text{ cm}^{-1}$ and $\sim 2854 \text{ cm}^{-1}$ (in Figure 3(b)) or in number of peaks at region 1750–1700 cm^{-1} (in Figure 3(c)). As mentioned in Section 2.3, spectrum of BCSO has two absorption peaks at $\sim 1743 \text{ cm}^{-1}$ and $\sim 1710 \text{ cm}^{-1}$ in the region of 1750-1700 cm^{-1} , while GSO has only one peak at 1743 cm^{-1} . These absorption peaks are mainly attributed to C=O carbonyl stretching vibration [34]. Another notable difference in the spectra of the two oils is that the absorption band at $\sim 1659 \text{ cm}^{-1}$ characteristic of thymoquinone was only observed in the spectrum of BCSO, but not in the spectrum of GSO.

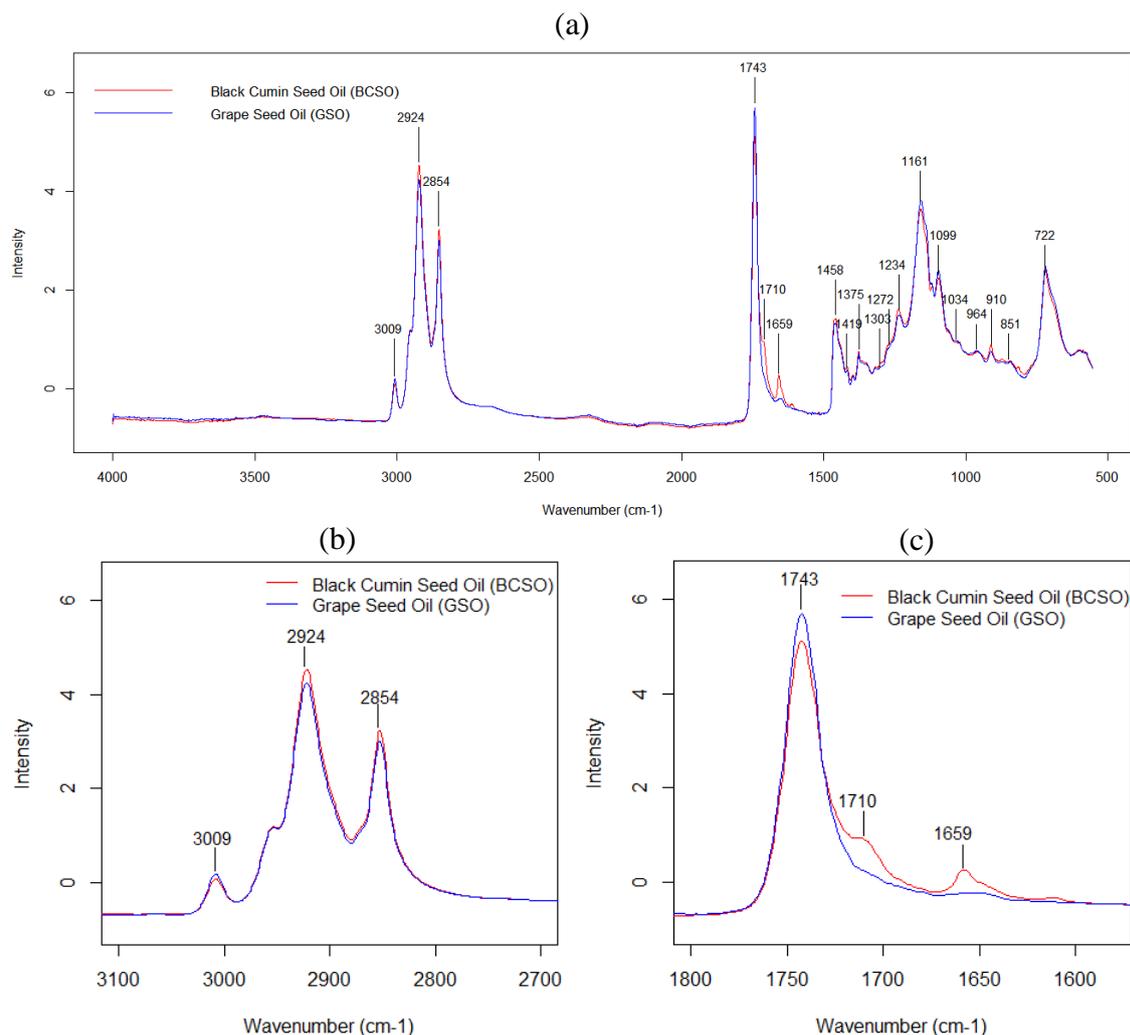


Figure 3. (a) Mean spectra from BCSO and GSO after standard pre-treatment labelled with major peaks of absorption bands. The magnified mean spectra in the regions of (b) 3100-2700 cm^{-1} and (c) 1800 -1600 cm^{-1} .

Table 1. Functional groups responsible for absorption band assignments of the ATR-FTIR spectrum of edible oil.

Wavenumber (cm^{-1})	Definition of spectral assignments
3009	C-H stretching of vinyl group
2924	CH_2 asymmetrical stretching, C-H stretching of aliphatic group
2854	CH_2 symmetrical stretching, C-H stretching of aliphatic group
1743, 1710	C=O carbonyl stretching vibration, characteristic to saturated fatty acid, lipid absorption.
1659	C=O stretching of thymoquinone
1458	CH_2 bending

1419	<i>Cis</i> =C-H bending
1375	CH ₃ symmetrical deformation
1303	Double links <i>Cis</i> unconjugated
1272	CH ₃ bending
1234	C-O stretch
1161	C-O stretch; CH ₂ bending
1099, 1034	C-O stretch, C-H bending group
964	<i>trans</i> CH=CH bending out of plane
910	C-H bending out of plane
851	CH ₂ wagging
722	CH=CH bending out of plane

3.2 Quantification of Adulteration by PLSR and Sparse PLSR

The PLSR model was built based on full spectrum region. The cross-validation results in quantification of BCSO were used to optimize k , the number of PLS components. In Table 2, using the first five PLS components, the optimal PLS regression model achieved the root mean squared error (RMSE) of 0.0206 using leave-out-one-section cross-validation. When the optimal model was used for prediction on the test set, it gave a 0.0175 prediction error. The sparse PLSR selected 211 out of 1149 spectral variables for the optimal model. The model can achieve a cross-validated RMSE of 0.0209 and a prediction error of 0.0173.

In the left panel of Fig. 4, the optimal PLSR model using the first five PLS components show high loadings in the regions corresponding to the spectral peaks. However, with all the spectral features used in the model, the loadings from the PLSR models look largely complex and thus the contribution of each wavelength to the linear diagnostic rule becomes less interpretable when it is related to the spectral features. In the right panel of Fig. 4, the sparse PLSR seems to shrink the PLSR loading coefficients with some group effect shown at certain regions. The coefficients of the sparse PLSR become much sparser and can be related easier to the spectral features as compared to the coefficients of PLSR model.

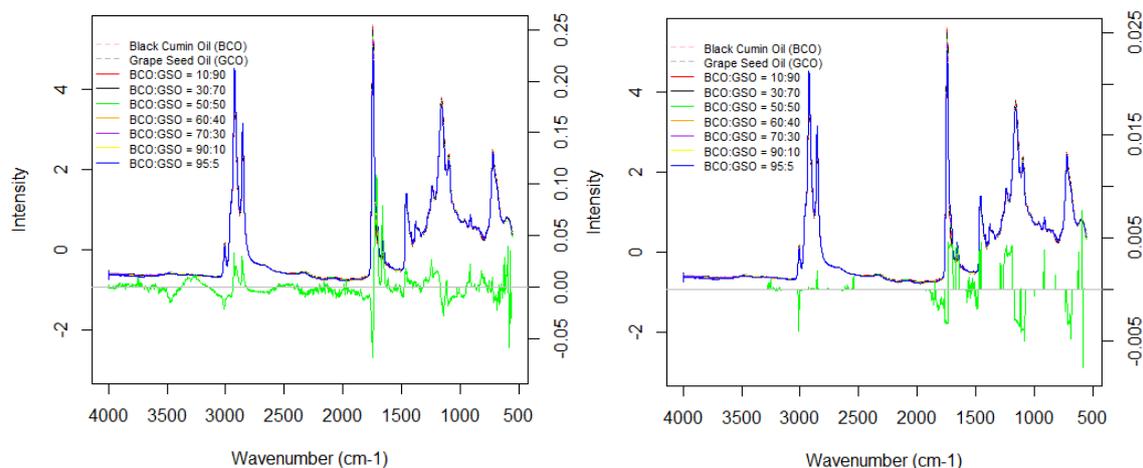


Figure 4. PLSR (left panel) and sparse PLSR (right panel) model coefficients for optimal tuning parameters. The model coefficients are shown in green, with mean spectra of mixture of BCSO and GSO in various mixing ratios superimposed.

3.3 Quantification of Adulteration by Elastic Net

The elastic net model aims to find the optimal parameters λ_1 and λ_2 for spectral variables selection and quantification of adulteration in BCSO. The model sparsity tends to increase as the parameter λ_1 increases. The value of $\lambda_1 = 0.6$ was chosen that gave the minimum cross-validated mean squared error. Figure 5 illustrates how a decreasing $\log(\lambda_2)$ leads to a decrease in cross-validated error for the elastic net model with the optimal λ_1 . It is observed that the cross-validated error is fairly flat over a range of $\log(\lambda_2)$ values less than -4.5 where shrinkage has been applied when the number of variables selected decreases with the increment of $\log(\lambda_2)$. In Figure 5, the λ_2 value with the minimum cross-validated error is indicated by the vertical dotted line. The optimal λ_2 value is indicated by the vertical dashed line where the largest value λ_2 lies such that the cross-validated error is no more than one standard error of the minimum. As shown in Table 2, the optimal elastic net model with $\lambda_1 = 0.6$ and $\log(\lambda_2) = -4.5$ yielded the RMSE of 0.0215 for quantifying GSO adulteration in BCSO by use of leave-out-one-section cross-validation. With 62 out of 1149 spectral variables chosen in the optimal model, the elastic net model gave a prediction error of 0.0230 for the separate test set.

The generated non-zero coefficients of the elastic net model suggested how those selected spectral regions contributed to accurate quantification of BCSO adulterated with GSO. In Figure 6, the absorption regions selected by the elastic net model are concentrated at $2900\text{-}2800\text{ cm}^{-1}$, $1750\text{-}1650\text{ cm}^{-1}$ and $1300\text{-}1050\text{ cm}^{-1}$.

For comparison purpose, the results of the ridge regression, as a specific case of elastic-net, are also provided in Table 2. Given that the error differences in ridge and elastic net are quite small, the two methods may be considered to give equivalently good prediction performance. The methods including all the spectral features sometimes may be

found to be more effective for prediction than those methods relying on feature selection. However, the model interpretation is quite challenging as shown in Figure 6 (right panel). Thus, for wavelength interpretation purpose, sparse models with informative features selected are often preferred.

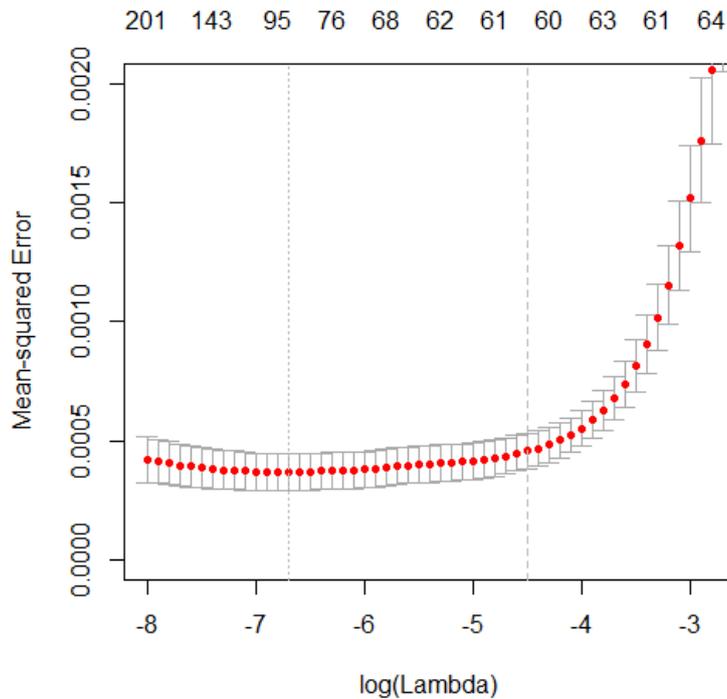


Figure 5. Cross-validated mean-squared error (\pm one standard deviation) for elastic net model against a sequence of tuning parameter values $\log(\lambda_2)$ with the number of non-zero coefficients for each parameter listed on the upper x-axis. The optimal value of λ_2 is indicated by the vertical dashed line. The λ_2 value with the minimum cross-validated error is indicated by the vertical dotted line.

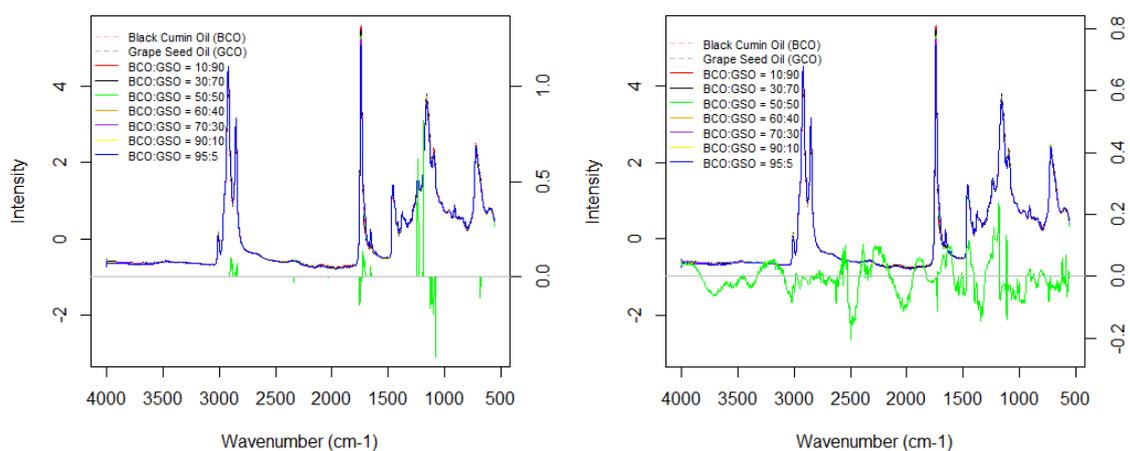


Figure 6. Elastic net (left panel) and Ridge (right panel) model coefficients for optimal tuning parameters. The model coefficients are shown in green, with mean spectra of mixture of BCSO and GSO in various mixing ratios superimposed.

Table 2. Results of the three penalized regression models for quantification of GSO adulterant in BCSO.

Method	Parameters	Non-zero features	Cross-validation RMSE	Prediction error
PLSR	$k = 5$	1149	0.0206	0.0175
Ridge	$\log(\lambda) = -1.6$	1149	0.0230	0.0214
Sparse PLSR	$k = 5, \lambda_1 = 0.9$	211	0.0209	0.0173
Elastic Net	$\lambda_1 = 0.6, \log(\lambda_2) = -4.5$	62	0.0215	0.0230
Group Lasso	$\log(\lambda) = -8.5$	61	0.0212	0.0233
Clustering Group Lasso	$\log(\lambda) = -9.2$	40	0.0205	0.0198

3.4 Quantification of Adulteration by Group Lasso

Although elastic net model tends to assign nearly identical coefficients to highly correlated predictor variables [19], its grouping effect is usually not apparent as expected [41,42]. To deal with this problem, ordinary group lasso model is used with prescribed partition of the predictors into groups to encourages sparsity at the group level.

The spectral variables were first partitioned into disjoint groups of equal size. Among those tried values for group size, a group size of 30 was chosen as it gave the best and stable cross-validation results in quantification of BCSO adulterated with GSO. The group lasso model seeks the optimal parameter λ that results in an effective variable selection in groups. Figure 7 illustrates that how a decreasing $\log(\lambda)$ leads to a decrease in mean squared error. While little change is observed in mean squared error for a range of $\log(\lambda)$ values less than -8.5, shrinkage has been applied when the number of variables selected in groups decreases with the increment of $\log(\lambda)$. In Figure 7, the vertical dashed line indicates the parameter $\log(\lambda) = -8.5$ of the optimal model whose error is no more than one standard error of the minimum. As shown in Table 2, the optimal group lasso model yielded a cross-validated RMSE of 0.0212 and a prediction error of 0.0233 for quantifying GSO adulterant in BCSO. With 61 out of 1149 spectral variables selected, the optimal group lasso model exhibited an equivalently good performance as compared to the elastic net model in terms of prediction accuracy and sparsity level.

The generated non-zero coefficients of the group lasso model indicated that informative spectral features selected by the group lasso model are mainly focused on the spectral region of 1750-1650 cm^{-1} as illustrated in Figure 8.

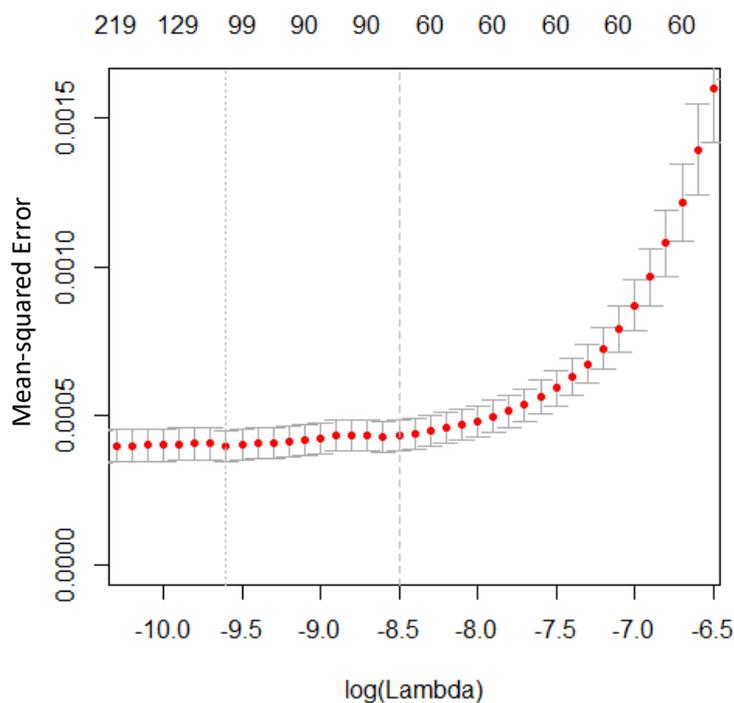


Figure 7. Cross-validated mean-squared error (\pm one standard deviation) for group lasso model against a sequence of tuning parameter values $\log(\lambda)$ with the number of non-zero coefficients for each parameter listed on the upper x-axis. The optimal λ and the λ with the minimum cross-validated error are indicated by the vertical dashed line and the dotted line, respectively.

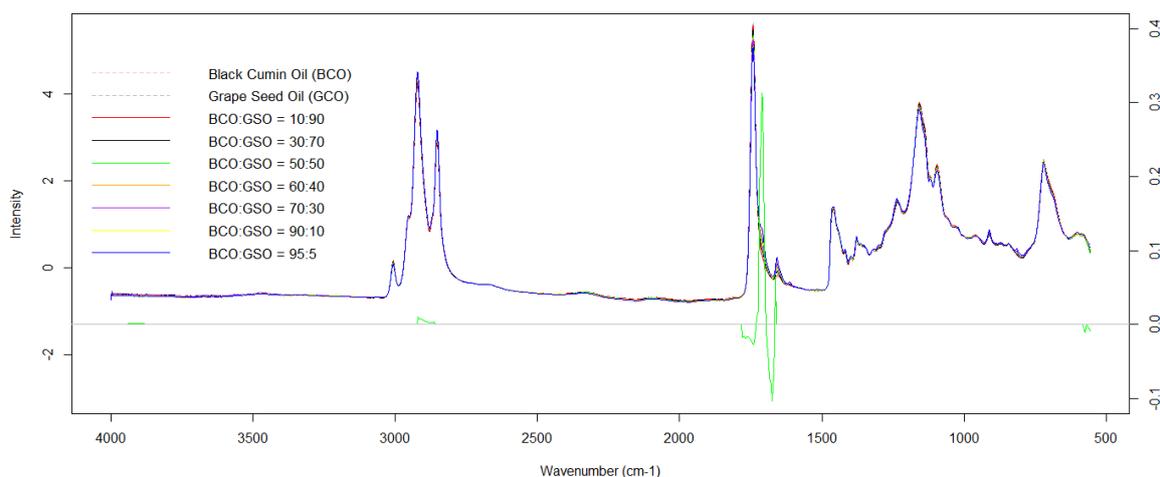


Figure 8. Group lasso model coefficients for optimal tuning parameter. The group lasso coefficients are shown in green, with mean spectra of mixture of BCSO and GSO in various mixing ratios superimposed.

3.5 Quantification of Adulteration by Clustering Group Lasso

Ordinary group lasso model assumes that the groups are predefined. When it is not obvious how to predefine the groups, the clustering group lasso model identifies the groups based on the spectral correlation structure.

The spectral variables were first grouped by hierarchical clustering and the number of 70 clusters was chosen through experiments as it produced good results in quantification of BCSO adulteration when it is combined with the group lasso model and the clustering is stable via cross-validation. Figure 9 illustrates that how a decreasing $\log(\lambda)$ results in a decrease in mean squared error. When little change is observed in mean squared error over a range of $\log(\lambda)$ less than -9.2, shrinkage has been applied on the process and the vertical dashed line locates the parameter $\log(\lambda) = -9.2$ of the optimal model whose error is no more than one standard error of the minimum. As given in Table 2, the optimal clustering group lasso model yielded a cross-validated RMSE of 0.0205 and a prediction error of 0.0198 for quantifying GSO adulterant in BCSO.

If the model improvement in cross-validated or prediction accuracy is not apparent, the improvement in model interpretability and reducing model complexity is significant. With only 40 (out of 1149) spectral variables selected, the optimal clustering group lasso model is the most parsimonious while achieving an excellent prediction performance as compared to the ordinary group lasso and the other penalized models, which is of particular advantage for wavelength interpretation.

In Figure 10, the generated non-zero coefficients of the clustering group lasso model indicated that the informative spectral features for accurate quantification of BCSO are focused on spectral regions of 2900-2800 cm^{-1} and 1750-1650 cm^{-1} .

The superior interpretation performance of the clustering group lasso model compared with the other three penalized models may be explained by the fact that FTIR spectrum is composed of many broad and overlapping absorption bands corresponding mainly to overtones and combinations of fundamental vibrations in chemical bonds [43]. The informative spectral variables for explaining variability in response variable show high correlation either in the nearest neighbours or over certain distance. The predefined groups for ordinary group lasso are usually based on the groups of adjacent spectral variables. Other spectral features over long distant are hardly captured by such predefined groups. When underlying spectral correlation structure was captured by the clustering method as groups over either close or distant range, the coefficients of those highly correlated spectral variables within a group share same or similar association with the response. The model thus enabled to identify the informative spectral features with apparent grouping effect as shown in Figure 10, which are associated with different chemical constituents in BCSO and GCO as described in Section 3.6.

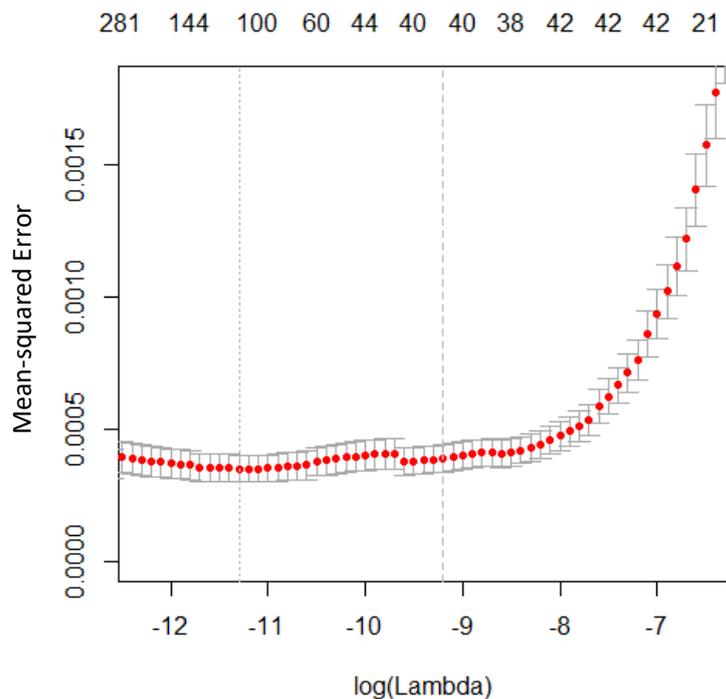


Figure 9. Cross-validated mean-squared error (\pm one standard deviation) for clustering group lasso model against a sequence of tuning parameter values $\log(\lambda)$ with the number of non-zero coefficients for each parameter listed on the upper x-axis. The optimal λ and the λ with the minimum cross-validated error are indicated by the vertical dashed line and the dotted line, respectively.

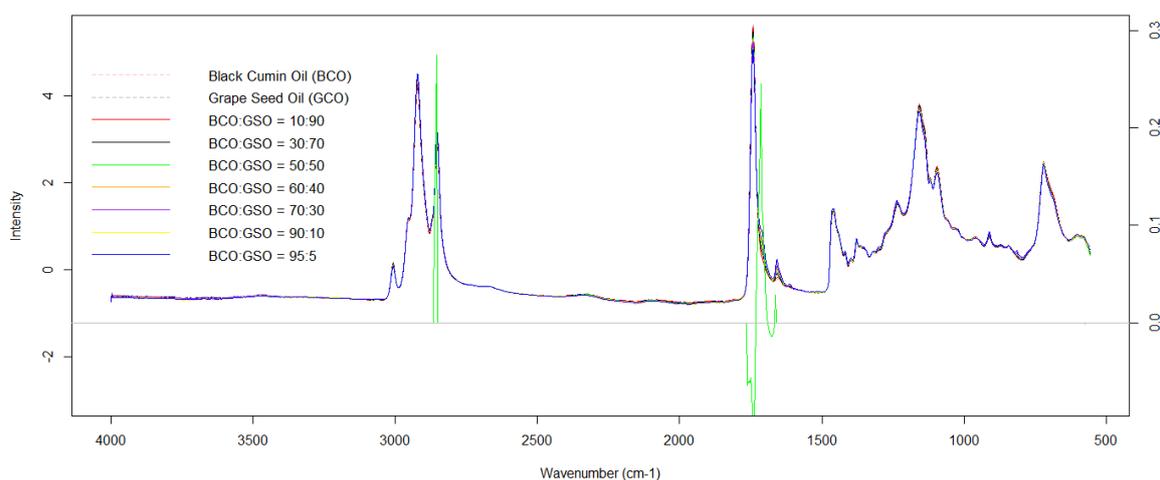


Figure 10. Clustering group lasso model coefficients for optimal tuning parameter. The clustering group lasso coefficients are shown in green, with mean spectra of mixture of BCSO and GSO in various mixing ratios superimposed.

3.6 Correlation between spectral absorption bands and chemical constituents of BCSO and its medicinal effect

Some chemical correlation can be found between spectral absorption bands and corresponding chemical constituents of BCSO. The spectral regions selected by the clustering group lasso model made a valuable contribution to interpreting the spectral features with respect to the major chemical constituents of BCSO and their distinct medicinal efficiency from a chemometric perspective. For instance, the selected regions in 1750-1650 cm^{-1} corresponding to C=O stretching vibration are characterized by the presence of high content of thymoquinone, which is a chief bioactive ingredient in BCSO and has been demonstrated with anticancer and analgesic effects. Another prominent absorption peak observed in selected region at 2900-2800 cm^{-1} due to CH_2 stretching vibration may indicate the presence of fatty acids like linoleic and oleic acids. It is known that linoleic acids in BCSO is very high compared to that in GCO, and it has been demonstrated to prevent oxidant.

The fact that the good quantification results were achieved by the clustering group lasso model, suggested intrinsic compositional differences between BCSO and GCO. The generated nonzero coefficients of the model provided a direct interpretation on how explicit spectral features contribute to the accurate authentication of BCSO. In particular, high coefficients at 1750-1650 cm^{-1} indicated that the spectra of BCSO and GCO may show different intensity or different features in this region, which may be associated with different content levels of thymoquinone in BCSO and GCO. It is observed that the spectra with high percentage of BCSO show two absorption peaks at $\sim 1743 \text{ cm}^{-1}$ and $\sim 1710 \text{ cm}^{-1}$ as mentioned in Section 2.3 and Section 3.1, while the spectra with low percentage of BCSO has only one peak at $\sim 1743 \text{ cm}^{-1}$ in the selected region.

In this study, though the chemical assignments of specific spectral absorption bands as given in Table 1 are not necessary for the data-driven feature selection, they are valuable for model interpretation regarding the important spectral features. In the case when the priori chemical knowledge of spectral features is not available, other methods may be considered to guide the understanding of the feature importance, such as stability selection method [44] based on subsampling in combination with regularized regression, and feature importance estimates derived from random forest [45] in order to perform robust feature selection.

4. Conclusion

BCSO constitutes valuable essential oils and fatty acids, and is thus one of the high value oils in fats and oils industry due to its nutritional applications and its beneficial effects on human health. Due to its high price value in market, BCSO is often a target to be adulterated with other cheaper oils. However, it is difficult to identify adulterated BCSO products through physical appearance. Therefore, it is essentially important to quantify BCSO adulteration quickly and reliably with the aim of BCSO authentication, quality assurance and scientific understanding of the constituents of BCSO.

In this study, the group lasso models applied on ATR-FTIR spectroscopy proved to be a powerful tool to quantify GSO adulterant in BCSO. The integration of ATR-FTIR technique and novel multivariate statistical methods provides a nondestructive, rapid and economic approach for direct determination of adulteration with subtle compositional differences among samples. Using a small subset of clustered spectral variables, the clustering group lasso model has achieved a cross-validated RMSE of 0.0205 for quantifying BCSO adulteration, which is considered to be rather small as compared to the actual percentage of BCSO in the mixed oil. Thus the model generally can quantify BCSO adulteration with high accuracy and can accurately quantify GSO adulterant at levels as low as 5%. The good performance was maintained with prediction error of 0.0198 on a test set. With spectral variables grouped in different ways, the two group lasso models explored in the paper allowed an automatic selection of spectral variables in groups that are most relevant to quantification of BCSO adulterated with GSO. The ordinary group lasso assumes groups known a priori. In literature, grouping information is often introduced into a model as a priori utilizing scientifically meaningful knowledge [27]. However, such grouping information is often not available. The two-stage clustering group lasso model can solve this problem by first inferring clusters (groups) of variables from the data followed by shrinking and selecting the identified groups of variables through group lasso method. When the high correlation structure of spectral variables and the underlying group features were both incorporated into the model, it encouraged correlated spectral variables within a group to have a shared association with the response. The grouping effect of the model is then apparent relative to the elastic net model and other penalized models. The two-stage clustering group lasso model is also flexible as different clustering approaches can be used to capture the underlying group structure so as to improve model performance. Other methods may be considered to cluster variables for use in group lasso, such as canonical correlation based clustering to address the problem of near linear dependence among variables [46], and tree-harvesting [47] which uses supervised learning methods to identify groups of variables.

With a small number of informative spectral variables selected, the clustering group lasso model achieved an equivalently excellent performance in both cross-validation and prediction accuracy, as compared to the elastic net model and the ordinary group lasso model for quantification of BCSO adulteration. The model was also compared with the widely used sparse PLSR. Though the sparse PLSR model shows comparable results for predicting percentage of pureness with some group effect, its penalty is imposed on each component independently and it leads to a relatively large number of variables and is less interpretable as compared to the usual penalized methods which provide a global selection of variables instead of a selection dimension by dimension. As a parsimonious model, the clustering group lasso model is of particular advantage for wavelength interpretation.

The most essential contribution of the clustering group lasso model lies in that the spectral variables selected in groups for quantifying adulteration are found to show a good link between the informative spectral features and the chemical constituents of BCSO. This implies that the accurate authentication of BCSO may be directly associated with the high-level contents of thymoquinone and fatty acids in BCSO. Thymoquinone as a main active constituent in BCSO is believed to be responsible for beneficial effects on human health

including antioxidant. The findings are novel and important in view of chemometrics. The quantitative interpretation of the results provides scientific evidence for many attributed health benefits and anticancer properties of BCSO. Based on these findings, it could be useful to further develop the agents derived from BCSO in modern medicine [48].

Considering a substantial reduction in number of spectral features and its simple regression coefficient profile, this clustering group lasso model may have regulatory and commercial potential for a robust and fast authentication of raw materials of BCSO in real-time platform, and can be widely applied for discriminating many other raw materials of traditional medicinal oils and food authentication. In further research, particular attention would be given to those identified spectral regions related to their bioactive chemical constituents in traditional medicinal oil.

Due to the limited resource of samples and the fact that conducting experiments of mixing samples are time-consuming, in this study only single batch of commercial oil bottles were used for estimating the purity of BCSO as a proof of concept. The results and chemical composition interpretations are based on the limited number of samples originated from the same batch and thus sharing the similar characteristics. This study is therefore considered as a first step feasibility study for the application concerned. In our future work, the clustering group lasso models will be further explored on a larger number of oil samples from different batches and different geographic origins, so as to give a better estimation of the expected prediction error in the real-life application.

Though the focus of this study is on a quantification problem, in real life practice, different types of oil adulterants may be considered for detection in adulterated oil samples. One-class-classification problem [49] can be explored for discrimination of pureness from adulteration before predicting the percentage of pureness. In future, our study will be extended to a detection problem, and penalized discrimination and classification methods will be further explored for detecting adulteration.

Acknowledgement

The authors would like to thank the editor and the anonymous reviewers for their valuable suggestions and comments that have helped us to improve the manuscript. This research is sponsored by Academic Research Funds (AcRF: RI 6/14 ZY) of National Institute of Education, Nanyang Technological University, Singapore

References

- [1] B. Amin, H. Hosseinzadeh, Black cumin (*Nigella sativa*) and its active constituent, Thymoquinone: An overview on the Analgesic and Anti-inflammatory effects, *Planta Med.* 82 (2016) 8-16.
- [2] A. Al-Hader, A. Aqel, Z. Hassan, Hypoglycemic effects of the volatile oil of *Nigella sativa* seeds, *International Journal of Pharmacology* 31 (1993) 96-100.
- [3] M.S. Al-Ghamdi, The anti-inflammatory, analgesic and antipyretic activity of *Nigella sativa*, *Journal of Ethnopharmacology* 76 (2001) 45-48.

- [4] M.A. Khan, Chemical composition and medicinal properties of *Nigella sativa* Linn, *Inflammopharmacol* 7 (1999) 15-35.
- [5] J. B. Rossell, B. King, M. J. Downes, Detection of adulteration, *Journal of the American Oil Chemists' Society* 60 (1983) 333–339.
- [6] A. K. Shukla, A. K. Dixit, R. P. Singh, Detection of adulteration in edible oils, *Journal of Oleo Science* 6 (2005) 317–324.
- [7] J. Moros, S. Garrigues, M. de la Guardia, Vibrational spectroscopy provides a green tool for multi-component analysis, *Trends in Analytical Chemistry* 29(7) (2010) 578–591.
- [8] Y.H. Lai, Y.N. Ni, S. Kokot, Classification of raw and roasted semen cassia samples with the use of Fourier transform infrared fingerprints and least squares support vector machines, *Appl. Spectrosc.* 64 (2010) 649-656.
- [9] M. Kiralana, G. Çalikb, S. Kiralana, A. Özaydinc, G. Özkand, M. F. Ramadane, Stability and volatile oxidation compounds of grape seed, flax seed and black cumin seed cold-pressed oils as affected by thermal oxidation, *Grasas Aceites* 70 (1) (2019) e295.
- [10] A.F. Nurrulhidayah, Y.B. Che Man, H.A. Al-Kahtani, A. Rohman, Application of FTIR spectroscopy coupled with chemometrics for authentication of *Nigella sativa* seed oil, *Spectroscopy* 25 (2011) 243–250.
- [11] D.M. Hawkins, The problem of overfitting, *Journal of chemical information and computer sciences* 44 (2004) 1-12.
- [12] Y. Zhu, T.L. Tan, Penalized discrimination analysis for the detection of wild grown and cultivated *Ganoderma lucidum* using fourier transform infrared spectroscopy, *J. Spectrochim. Acta Part A Mol. Biomol. Spectrosc.* 159 (2016) 68–77.
- [13] T. Næs, H. Martens, Principal component regression in NIR analysis: viewpoints, background details and selection of components, *J. Chemom.* 2 (2) (1988) 155–167.
- [14] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, *Anal. Chim. acta* 185 (1986) 1–17.
- [15] S. Wold, Personal memories of the early PLS development, *Chemom. Intell. Lab. Syst.* 58 (2) (2001) 83–84.
- [16] H. Zou, T. Hastie, R. Tibshirani, Sparse principal component analysis, *Journal of Computational and Graphical Statistics* 15(2) (2006) 265–286.
- [17] D. Rossouw, C. Robert-Granié, P. Besse. A sparse pls for variable selection when integrating omics data, *Genetics and Molecular Biology* 7 (1) (2008) 35.

- [18] H. Chun, S. Keles, Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (1) (2010) 3–25.
- [19] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B* 76 (2005) 301–320.
- [20] M. Yuan, Y. Lin, Model Selection and Estimation in Regression with Grouped Variables, *Journal of the Royal Statistical Society Series B* 68 (2006) 49-67.
- [21] B. Stuart, Biological Applications of Infrared Spectroscopy, in: *Analytical Chemistry of Open Learning* (115), John Wiley & Sons, Chichester, 1997.
- [22] T. Næs, T. Isaksson, T. Fearn, T. Davies, *A User-Friendly Guide to Multivariate Calibration and Classification*, NIR Publications, Chichester, UK, 2002.
- [23] A. Savitzky, M.J.E. Golay, Smoothing and Differentiation of Data by Simplified Least-Squares Procedures, *Analytical Chemistry* 36 (1964) 1627-1639.
- [24] R.J. Barnes, M.S. Dhanoa, S.J. Lister, Standard Normal Variate Transformation and De-Trending of Near-Infrared Diffuse Reflectance Spectra, *Applied Spectroscopy* 43 (1989) 772-777.
- [25] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. R. Stat. Soc. B* 58 (1996) 267–288.
- [26] A.E. Hoerl, R. Kennard, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics* 12 (1970) 55–67.
- [27] J. Huang, P. Breheny, S. Ma, A Selective Review of Group Selection in High-Dimensional Models, *Statistical Science* 27 (4) (2012) 481-499.
- [28] E. Vigneau, E.M. Qannari, Clustering of variables around latent component, *Comm. Stat. Simul. Comput.* 32 (2003) 1131–1150.
- [29] T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, New York, 2009.
- [30] R.A. Johnson, D.W. Wichern, *Applied Multivariate Statistical Analysis*, Prentice Hall, 2002.
- [31] J. Camacho, A.K. Smilde, E. Saccenti, J. A. Westerhuis, Rasmus Bro, All sparse PCA models are wrong, but some are useful. Part II: Limitations and problems of deflation, *Chemometrics and Intelligent Laboratory Systems* 208 (2021) 104212.
- [32] M. Chavent, B. Liquet, V. Kuentz, J. Saracco, ClustOfVar: An R Package for the Clustering of Variables, *Journal of Statistical Software* 50 (2012) 1-16.

- [33] R. Core Team, R: a Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2012.
- [34] S.J. Mohammed, H.H.H. Amin, S.B. Aziz, A.M. Sha, S. Hassan, J.M.A. Aziz, H.S. Rahman, Structural characterization, antimicrobial activity, and in vitro cytotoxicity effect of black seed oil, *Evidence-Based Complementary and Alternative Medicine* 2019 (2019) 6515671.
- [35] D.L. Pavia, G.M. Lampman, G.S. Kriz-jr, Introduction to spectroscopy: a guide for students of organic chemistry, 4th Ed, Thomson Learning Inc.: London, UK, 2001.
- [36] G. Singh, P. Marimuthu, C.S. de Heluani, C. Catalan, Chemical constituents and antimicrobial and antioxidant potentials of essential oil and acetone extract of *Nigella sativa* seeds, *Journal of the Science of Food and Agriculture* 85(13) (2005) 2297-2306.
- [37] B.H. Ali, G. Blunden, Pharmacological and toxicological properties of *Nigella sativa*, *Phytotherapy Research* 17(4) (2003) 299-305.
- [38] H. Lutterodt, M. Luther, M. Slavin, J.J. Yin, J. Parry, J.M. Gao, L.L. Yu, Fatty acid profile, thymoquinone content, oxidative stability, and antioxidant properties of cold-pressed black cumin seed oils, *LWT-Food Science and Technology* 43(9) (2010) 1409-1413.
- [39] C.C.Woo, A.P. Kumar, G.Sethi, K.H.B. Tan, Thymoquinone: potential cure for inflammatory disorders and cancer, *Biochemical Pharmacology* 83(4) (2012) 443-451.
- [40] S. Pagola, A. Benavente, A. Raschi, E. Romano, M. A. A. Molina, P. W. Stephens, Crystal Structure Determination of Thymoquinone by High-Resolution X-ray Powder Diffraction, *AAPS PharmSciTech* 5(2) (2004) 28.
- [41] Y. Zhu, T. L. Tan, W.K. Cheang, Penalized logistic regression for classification and feature selection with its application to detection of two official species of *Ganoderma*, *Chemometrics and Intelligent Laboratory Systems* 171 (2017) 55-64.
- [42] J. Xie, L. Zeng (2010) Group Variable Selection Methods and Their Applications in Analysis of Genomic Data. In: Feng J., Fu W., Sun F. (Eds.), *Frontiers in Computational and Systems Biology*, vol 15, Computational Biology, Springer, London, 2010, pp. 231-248.
- [43] B.G. Osborne, Near-infrared spectroscopy in food analysis, in: *Encyclopedia of Analytical Chemistry*, John Wiley & Sons, New York, 2006.
- [44] N. Meinshausen, P. Bühlmann, Stability selection, *J. Royal Stat. Soc. Ser. B (Statistical Methodol)*, 72 (2010) 417-473.
- [45] G. Louppe, L. Wehenkel, A. Suter, P. Geurts, Understanding variable importances in forests of randomized trees. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, K.Q. Weinberger, (eds.) *Advances in Neural Information Processing Systems*, vol 26, Curran Associates, Inc., NY, United States, 2013, pp. 431-439.

- [46] P. Buhlmann, P. Rutimann, S. van de Geer, C.H. Zhang, Correlated variables in regression: clustering and sparse estimation, *Journal of Statistical Planning and Inference* 143 (2012) 1835-1871.
- [47] T. Hastie, R. Tibshirani, D. Botstein, P. Brown, Supervised harvesting of expression trees, *Genome Biology* 2 (2001) 1-12.
- [48] S. Padhye, S. Banerjee, A. Ahmad, R. Mohammad, F.H. Sarkar, From here to eternity -- The secret of Pharaohs: Therapeutic potential of black cumin seeds and beyond, *Cancer Therapy* 6(b) (2008) 495-510.
- [49] O.Y. Rodionova, P. Oliveri, A.L. Pomerantsev, Rigorous and compliant approaches to one-class classification, *Chemometrics and Intelligent Laboratory Systems* 159 (2016) 89-96.