Anal Bioanal Chem (2017) 409:6699–6708
DOI 10.1007/s00216-017-0628-8

RESEARCH PAPER

Estimation of start and stop numbers for cluster resolution feature selection algorithm: an empirical approach using null distribution analysis of Fisher ratios

Lawrence A. Adutwum[1] · A. Paulina de la Mata[1] · Heather D. Bean[2] · Jane E. Hill[3] · James J. Harynuk[1]

Received: 25 July 2017 / Revised: 29 August 2017 / Accepted: 6 September 2017 / Published online: 29 September 2017

Abstract Cluster resolution feature selection (CR-FS) is a hybrid feature selection algorithm which involves the evaluation of ranked variables via sequential backward elimination (SBE) and sequential forward selection (SFS). The implementation of CR-FS requires two main inputs, namely, start and stop number. The start number is the number of the highly ranked variables for the SBE while the stop number is the point at which the search for additional features during the SFS stage is halted. The setting of these critical parameters has always relied on trial and error which introduced subjectivity in the results obtained. The start and stop numbers are known to vary with each dataset. Drawing inspiration from overlapping coefficients, a method for comparing two probability density functions, empirical equations toward the estimation of start and stop number for a dataset were developed. All of the parameters in the empirical equations are obtained from the comparisons of the two probability density functions except the constant termed *d*. The equations were optimized using three real-world datasets. The optimum range of *d* was determined to be 0.48 to 0.57. An implementation of CR-FS using two new datasets demonstrated the validity of this approach. Partial least squares discriminant analysis (PLS-DA) model prediction accuracies increased from 90 and 96 to 100% for both datasets using start and stop numbers calculated with this approach. Additionally, there was a twofold increase in the explained variance captured in the first two principal components.

Keywords Feature selection · Cluster resolution · Classification · Chemometrics · Overlapping coefficient · Fisher ratio

Introduction

Modern instrumental techniques in analytical chemistry generate huge amounts of data. This is most likely due to advancements in data acquisition technologies over the years [1, 2]. Analytical separation techniques such as gas chromatography (GC)/liquid chromatography (LC) with mass spectrometry (MS), and spectroscopic techniques such as nuclear magnetic resonance (NMR) and near infrared (NIR), are capable of generating several hundreds or even thousands of variables to describe an individual sample. Challenges that come with high number of variables are storage, distribution, and analysis [3]. Chemometric techniques combine statistical and computational methods to extract useful information from complex chemical data and have become very useful for handling chromatographic and spectroscopic data [4, 5]. Chemometric techniques have been applied to analytical chemistry data in several areas including pharmaceutical [6], petrochemical [7–9], forensics [10–14], food-omics [15–18], metabolomics [19–22], and biomarker discovery [23–25], amongst others.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00216-017-0628-8) contains supplementary material, which is available to authorized users.

✉ James J. Harynuk
james.harynuk@ualberta.ca

[1] Department of Chemistry, University of Alberta, 11227 Saskatchewan Drive NW, Edmonton, Alberta T6G 2G2, Canada

[2] School of Life Sciences, Arizona State University, 427 E Tyler Mall, Tempe, AZ 85287, USA

[3] Thayer School of Engineering, Dartmouth College, 14 Engineering Drive, Hanover, NH 03755, USA

Data collected from most scientific experiments includes portions generated by the chemical information in addition to artifacts and noise resulting from stochastic variations in experimental conditions and instruments. Artifacts, noise, or irrelevant variables in the data negatively impact chemometric models as the mathematics of the model must attempt to account for these meaningless variations [26, 27]. Prior to chemometric analysis, data are usually subjected to a series of mathematical transformations, termed preprocessing, aimed at reducing the influence of irrelevant signals on the data [28]. Signal smoothing, baseline drift correction, scaling, and centering are amongst the common preprocessing methods applied to data to be analyzed [29–31].

Feature selection, also referred to as variable selection, is the process of selecting a subset of the variables in the data which are relevant to the intended model [26, 27]. Features are considered relevant to a task or a research problem if alteration of their values in a sample alters the output of a target function [32–34]. Feature selection has been widely researched and many algorithms have been developed [26, 27, 34–38]. It has been established in both theory and practice that feature selection effectively enhances or increases predictive accuracy and aids in obtaining more parsimonious models [36, 39].

Feature selection algorithms are classified into three main groups, namely, filters, embedded, and wrapper methods [26]. Filter methods, otherwise referred to as ranking methods, compute a metric that is used to organize features according to their relevance/importance [26, 27, 38]. Variable relevance is estimated independently of the classifier. This makes filters computationally fast but the possibility of evaluating feature interdependencies is obviated. Some of the popular filter methods employed are the RELIEF algorithm, the FOCUS algorithm, and correlation-based filters [35, 38]. For wrapper methods, the relevance of variables is evaluated with the aid of an induction algorithm to assess the true importance, i.e., feature selection is wrapped around the induction algorithm [26, 27, 34, 40, 41]. Thus, the induction algorithm is used as part of the feature selection process as an evaluation function to help estimate the worth of each variable or a set of variables. The variable subspace can be searched randomly or via a sequential backward elimination (SBE) or sequential forward selection (SFS). Wrapper methods generally perform better than filter methods but are computationally expensive. There is an associated risk of overfitting as the feature selection is tuned to the specific dataset by the induction algorithm [38]. In embedded methods, variable importance is obtained from the induction/learning algorithm [26, 27]. Typical examples are the weighing vector of support vector machines (SVM), variable importance to projection (VIP), and selectivity ratio (SR) in PLS-DA [42–44].

In an earlier study, we reported an automated hybrid feature selection algorithm termed cluster resolution feature selection (CR-FS) [45, 46]. This is based on a model parameter termed cluster resolution (CR), which measures the separations between clusters of samples in different classes in a reduced dimensionality space. Irrespective of the number of classes, a single quality parameter bounded by 0 and 1 is computed to estimate the overall model quality. This makes it useful for the simultaneous optimization of multiclass problems. CR-FS combines a filter method such as Fisher ratio (F-ratio) from analysis of variance (ANOVA) to initially order variables [47]. It then searches the variable subspace via SBE and SFS to find variables that improve CR. During the SBE, an initial population of highly ranked variables are used to generate a PCA model. The number of highly ranked variables used is termed the start number. SBE then proceeds, testing each of the initial variables in order from the lowest-ranked to the highest-ranked with CR being evaluated at each step. Variables whose elimination does not lead to a deterioration of the model are discarded. In the SFS, variables that were not tested in the SBE are evaluated in order of decreasing F-ratio, i.e., reducing relevance. Features whose inclusion do not improve the CR are eliminated. The total number of variables evaluated in both the SBE and SFS is termed the stop number. CR-FS has been successfully applied to various types of problems and has always led to improved model prediction sensitivity, specificity, and accuracy [11, 45, 46, 48–50]. Algorithms based solely on SBE are known to be greedy while those based on SFS are likely to suffer a nesting problem [27, 33]. Nesting implies that once a variable is added through SFS or removed through SBE, it is permanently included in (or excluded from) the final model. This makes the start and stop number for SBE and SFS a critical parameter for CR-FS. In all previous studies, setting the start and stop numbers has been a matter of trial and error, relying on the experience of the user. This introduces subjectivity in the feature selection process, slows down the process, and prevents the true automation of CR-FS. Where fewer variables exist for the dataset, all of them can be evaluated but this may not be practical when several thousands or millions of variables exist for each sample, as is the case with raw chromatographic data. Previous experience with CR-FS demonstrates that the start and stop numbers can influence which features are retained. This makes it very important to find an objective and unbiased approach for choosing the start and stop number for CR-FS.

When class labels are uncoupled from a dataset and reassigned randomly, the F-ratios calculated from these misclassified datasets are termed null F-ratios. This is because they are generated under a distribution where the new means vary from the true class means. Comparison of the distribution of null and true F-ratios (obtained with correct class assignments) can provide useful information in determining the limit beyond which a true F-ratio may not be as informative. This technique has been applied to the analysis of comprehensive two-dimensional gas chromatographic data to reduce false positive rates [51]. Overlapping coefficient (OVL) also known

as Weitzman's measure is a measure of similarity (and for that matter dissimilarity) between two probability distributions represented by continuous probability density functions [52, 53]. It was first used by Weitzman to determine the degree of overlap of income distributions between families in the USA [52]. Even though other similarity measures such as Matusita's and Morisita's are available, OVL is preferred due to its simplicity and naturalness [53–55]. OVL compares the density functions for two probability distributions and relates the similarity to the overlapping regions of the area under the two density functions [53].

In this study, we draw inspiration from OVL to find the dissimilarities between true and null F-ratios from a dataset. The degree of dissimilarity is used as a guide to determine the number of variables that have a higher probability of being from the true F-ratios. Parameters obtained from the two density functions are used to devise empirical equations to estimate the start and stop numbers for CR-FS. The empirical equations are then tested with real data with the aim of finding the optimal parameters. It is hoped that this will eliminate the subjective and trial and error approach to the implementation of CR-FS.

Theory

The start and stop number for CR-FS are critical parameters. The availability of an empirical equation to estimate these parameters will make CR-FS fully automated. In the context of categorizing variables into whether they are truly relevant or not, a comparison of the probability density functions of the true F-ratio and the null F-ratio can be very informative. F-ratios with higher probability density in the true distribution relative to the null implies they are relatively more likely to come from the true distribution. Since the aim is to find F-ratio belonging to the true F-ratio, the focus is on the non-overlapping region of the two density functions.

True and null F-ratios

The F-ratio (f) is ratio of between class variance (σ^2_{bc}) to within class variance $\left(\sigma^2_{wc}\right)$ and it is calculated as shown in Eq. 1a, 1b, and 1c [47].

$$f = \frac{\sigma^2_{bc}}{\sigma^2_{wc}} \tag{1a}$$

$$\sigma^2_{bc} = \frac{\sum n_i \left(\bar{x}_i - \bar{x}\right)^2}{(K-1)} \tag{1b}$$

$$\sigma^2_{wc} = \frac{\sum_{i=1}^{K} \sum_{j=1}^{n_i} \left(x_{ij} - \bar{x}_i\right)^2}{(N-K)} \tag{1c}$$

where n_i is the number of variables in the ith class/group, $\bar{x}_i$ is the mean of the ith class, $\bar{x}$ is the data mean, $\bar{x}_{ij}$ is the jth observation in the ith out of K class/group, and N is the sample size.

True F-ratios (f_{TRUE}) are calculated from a subset of the training set data. Null F-ratios (f_{NULL}) are also calculated from the same subset after swapping the class assignment of approximately 10–15% of the data in each class, chosen at random. The rationale behind choosing 10–15% of the data to swap is that this introduces enough errors into the data to significantly affect the variance of each class without having too large an effect on the means of the classes.

Proposal of empirical equation for estimating start (n_{ST}) and stop numbers (n_{SP})

f_{TRUE} and f_{NULL} are fitted to a selection of continuous probability density functions (PDFs). Density functions for continuous probability distributions were evaluated. These included Weibull, chi-square, inverse Gaussian, log-normal, logistic, log-logistic, Gumbel, and Frechet. To determine the optimum PDF, the Akaike information criteria (AIC) was used. After fitting the data, the AIC of all the PDFs were estimated. Based on the lowest AIC, the optimal PDF is selected. A detailed review of AIC can be found here [56, 57]. If f_{T} and f_{N} are the optimal PDFs for f_{TRUE} and f_{NULL}, respectively, then a simultaneous plot of these two PDFs provides very useful information (Fig. 1 f_{T}—blue line, f_{N}—red line). The point of intersection (b) of the two PDFs can be determined by equating the two functions, i.e., $f_{\mathrm{T}}(b) = f_{\mathrm{N}}(b)$. The area under $f_{\mathrm{N}}(x)$, shaded in red, and $f_{\mathrm{T}}(x)$, where $x > b$, are estimated from Fig. 1. The area represented by the blue region is also determined. The area of the blue region relates to the cumulative density of $f_{\mathrm{T}}(x)$, where $b \le x \le k$, and k is the maximum f_{TRUE}. From the analysis of Fig. 1 and based on our experience, two empirical equations (2 and 3) were proposed to estimate the start number (n_{ST}) and stop number (n_{SP}), respectively.

$$n_{ST} = {n_{SP}}^{d} + (b \times f_T(b)) \tag{2}$$

$$n_{\mathrm{SP}} = \frac{\int_b^k f_T(x)dx - \int_b^m f_N(x)dx}{\int_b^k f_T(x)dx} \times C \tag{3}$$

where n_{ST} is the start number for SBE; n_{SP} is the stop number for SFS; f_{TRUE} are the true F-ratios; f_{NULL} are the null F-ratios; f_{T} is the optimal PDF for f_{TRUE}; f_{N} is the optimal PDF for f_{NULL}; k is the maximum f_{TRUE}; m is the maximum f_{NULL}; C is the number of variables in $f_{\mathrm{TRUE}} > b$; and d is a constant. With the exception of d, all parameters are obtained from the data.

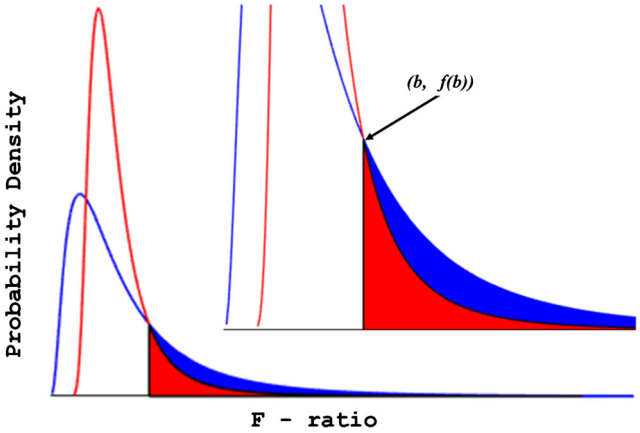

Fig. 1 A plot of probability density function for f_{TRUE}(f_T) and f_{NULL} (f_N). Red-shaded region represents area under the optimum density function of f_{NULL}, f_N with F-ratios greater than b. The sum of the blue and red regions represents area under the optimum density function for f_{TRUE}, f_T with F-ratios greater than b and $f_T(x) > f_N(x)$, where x are F-ratio > b and b is the F-ratio value at which they two density functions intersect

Chemometric analysis

Datasets

Five different datasets were used for this study. All the variables in each dataset were peak areas for LC-MS or cumulative peak areas of compounds (GC×GC-TOFMS) obtained using the corresponding commercial instrumental software. Dataset 1 (bac—bacteria) was obtained from the GC×GC-TOF MS analysis of the volatilome (i.e., volatile metabolites), of a suite of bacterial samples. The data consisted of 63 samples and 1673 variables to be classified as type 1 vs. type 2, having 35 and 28 samples, respectively. Dataset 2 (ucp—unwashed cotton polyester) was obtained from GC×GC-TOF MS analysis of volatile compounds extracted from worn cotton and polyester fabrics which have not been washed. The data consisted of 80 samples and 2781 variables. It was to be classified as unwashed cotton vs. unwashed polyester. Dataset 3 (wcp—washed cotton polyester) was obtained from GC×GC-TOF MS analysis of volatile compounds from worn cotton and polyester fabrics after they have been washed. It consisted of 80 samples and 2781 variables. It was to be classified as washed cotton vs. washed polyester. Dataset 4 (coff—coffee) was a peak table obtained from the LC-MS analysis of coffee samples. It consisted of 78 samples and 701 variables. It was to be classified as Arabica (Ara) vs. a mixture of Arabica and Robusta (Ara + Rob) coffee. Dataset 5 (cvp—cotton vs. polyester) was obtained from GC×GC-TOF MS analysis of volatile compounds. It consisted of 168 samples and 2781 variables. It was to be classified as cotton vs. polyester.

Detailed experimental conditions about datasets 2, 3, and 5 can be obtained from an earlier published work [49]. The details of the bacterial culture, metabolite collection and GC×GC-TOF MS for dataset 1 can be found in the Electronic Supplementary Material (ESM) section S1. Detail of sample preparation and analysis of dataset 4 can be found here [58]. Datasets 1, 2, and 3 were used to find the optimum value for the constant, d, in Eq. 3. Using the optimum value of d, the validity of the approach was tested with datasets 3 and 4.

Data importation and all computations were performed in Matlab® 2016b using in-house written algorithms. Chemometric models were constructed using PLS Toolbox 8.2.1 (Eigenvector Research Inc., Wenatchee, WA). All chemometric analyses were performed on 64-bit Windows 7 Enterprise running on a core i7—4790 K Intel processor and 32 GB RAM.

Estimation of the constant d and n_{SP}

Each of the three datasets (i.e., 1, 2, and 3) were split into two-thirds for training and one-third for external validation sets. Using half of the training set data, F-ratio analysis was performed as previously described. n_{ST} and n_{SP} were estimated from the probability density functions. n_{ST} is determined for a set of d values, as shown in Eq. 3, such that $0.05 \leq d \leq 0.95$. Using these n_{ST} and n_{SP} values, CR-FS was implemented on the entire training set data. This step was repeated ten times. During each iteration, a different subset of the training set data was used for F-ratio analysis and model optimization. Variables that were selected in at least six iterations were used for model evaluation. A PCA model and a PLS-DA model were constructed with the training set data using only the selected variables. The validation set data was projected into the PCA model and the validation set CR (cr_{max}) determined. The PLS-DA model prediction accuracy of the validation set was also determined. The product of the validation set CR (cr_{max}) and PLS-DA prediction accuracy was used as the objective parameter in determining the best value for d.

Results and discussion

CR-FS is a hybrid (filter and wrapper) feature selection algorithm that has been useful for improving classification accuracies of chemometric models. The two main parameters required by the algorithm are the n_{ST} and n_{SP}, for the SBE and SFS, respectively. The lack of a guidance as to the choice of these parameters introduces subjectivity and increases the feature selection time due to the trial and error nature of the optimization of these parameters. The aim of this study was to devise an empirical approach to the determination of n_{ST} and n_{SP}. This would eliminate subjectivity and allow for the true automation of the entire feature selection process. The n_{ST} and n_{SP} in an optimization with CR-FS varies with the dataset. It was also observed that the probability density of the F-ratio also varies with the dataset. Hence, it is possible to generalize

the n_{ST} and n_{SP} by connecting it to the PDFs of the F-ratios. Comparison of the PDFs obtained from f_{TRUE} and f_{NULL} was made using the concept of OVL to guide the proposal of an empirical equation (Eqs. 2 and 3).

The optimum density function, i.e., f_T and f_N, varies with the dataset. Hence, to determine the optimum PDF for f_{TRUE} and f_{NULL} for a dataset, several continuous density functions were tested. Amongst the distributions evaluated were Weibull, chi-square, inverse Gaussian, log-normal, logistic, log-logistic, Gumbel, and Frechet. Since the optimum PDF is unknown, determination of the AIC provided a means to evaluate the PDFs. AIC is a measure of relative quality of statistical models used to fit the same data [56, 57, 59, 60]. The use of AIC to determine the optimum PDF eliminates the risks associated with overfitting or underfitting the data. The two PDFs, i.e., f_T and f_N, do not necessarily have to be the same. Thus, irrespective of the density functions, a figure similar to Fig. 1 results.

All but one parameter, i.e., d, in the empirical equations can be obtained from the analysis of the density functions of f_{TRUE} and f_{NULL}. It can be deduced from Eq. 2 that a lower d value yields a smaller n_{ST} and the feature selection is dominated by SFS. A higher d value on the other hand yields a higher n_{SP} which makes CR-FS SBE-dominated. Three of the datasets were used to determine the optimum value of d in Eq. 2. In order to simultaneously compare the results of PCA and PLS-DA models, the product of the CR of the validation set data (cr_{max}) from PCA and model prediction accuracy (*acc*) of the PLS-DA, i.e., $cr_{max} \times$ acc, was used as the objective model quality parameter. A plot of the results is shown in Fig. 2. From Fig. 2a, it is obvious that lower d values seem to lead to better models; however, Fig. 2b shows the standard deviation of the $cr_{max} \times$ acc is higher at lower d values. During each iteration, a different subset of the training data was used for training and optimization as indicated earlier. If n_{ST} is too low, retained features tend to overfit the model to that specific subset. Hence when applied to the external validation set, high variability in prediction accuracies occurs. The standard deviation decreases as the n_{ST} is increased (Fig. 2b). As d increases beyond 0.65, the model prediction capability for all datasets starts to deteriorate. This is because at high d values, n_{ST} tends to be high (Eq. 2). CR-FS performed with n_{ST} are dominated by SBE. Since SBE is greedy, several variables that may not be highly relevant end up in the model and lead to poor prediction accuracies. Figure 3 shows a z-score (mean/σ) plot used to identify the region with good model prediction accuracies by lower deviations. The region for d such that $0.48 \le d \le 0.57$ tends to have a higher model prediction quality with an accompanied lower standard deviation. Thus, if CR-FS implemented with n_{ST} estimated with d between 0.48 and 0.57, a core number of features are retained, which leads to good predictions irrespective of the subsets of the training data used for optimization.

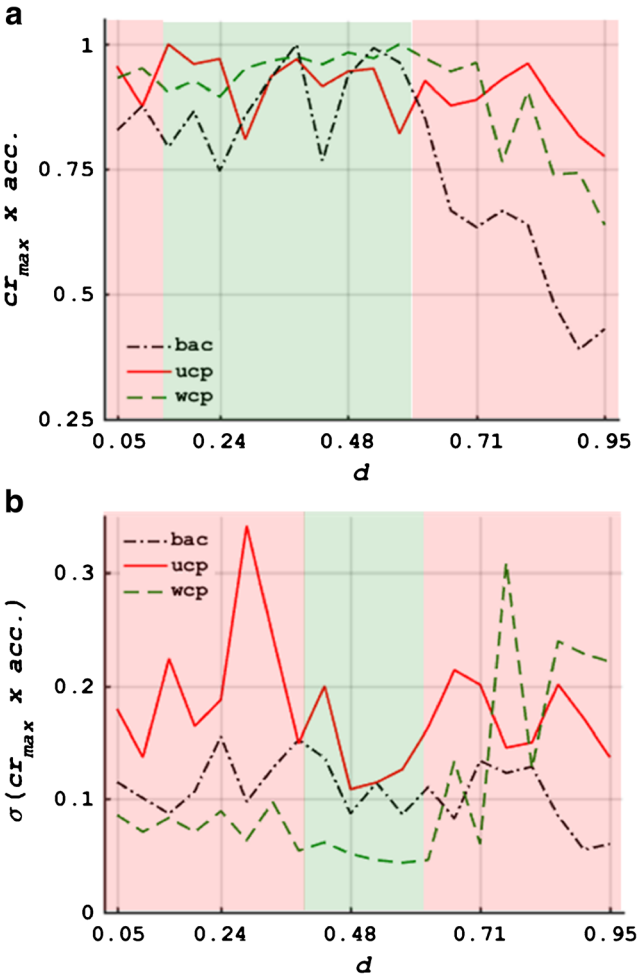

Fig. 2 A plot of model quality as a function of d ($0.05 \le d \le 0.95$). Model quality $cr_{max} \times$ acc vs. d is shown in **a** while **b** shows the standard deviation, $\sigma(cr_{max} \times acc)$ vs. d. Variation in d influences the n_{ST} for SBE according to Eq. 2. n_{SP} was determined by Eq. 3

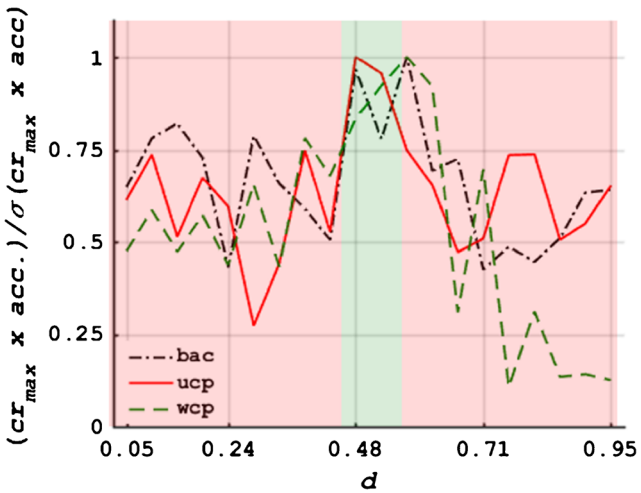

Fig. 3 A plot of the ratio of model quality parameter/σ as a function of d ($0.05 \le d \le 0.95$)

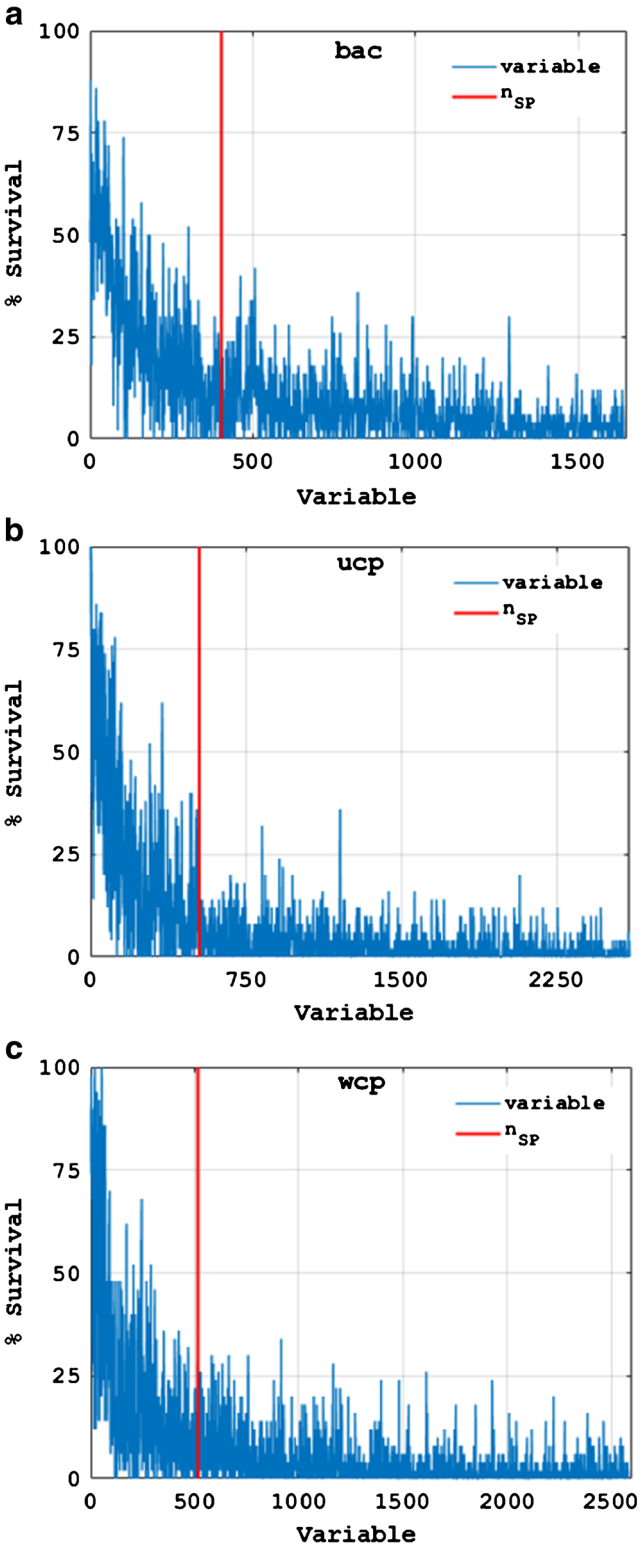

Fig. 4 Feature survival rate for all variables for **a** bac, **b** ucp, and **c** wcp. n_{SP} was estimated from the optimum d value; all variables were evaluated

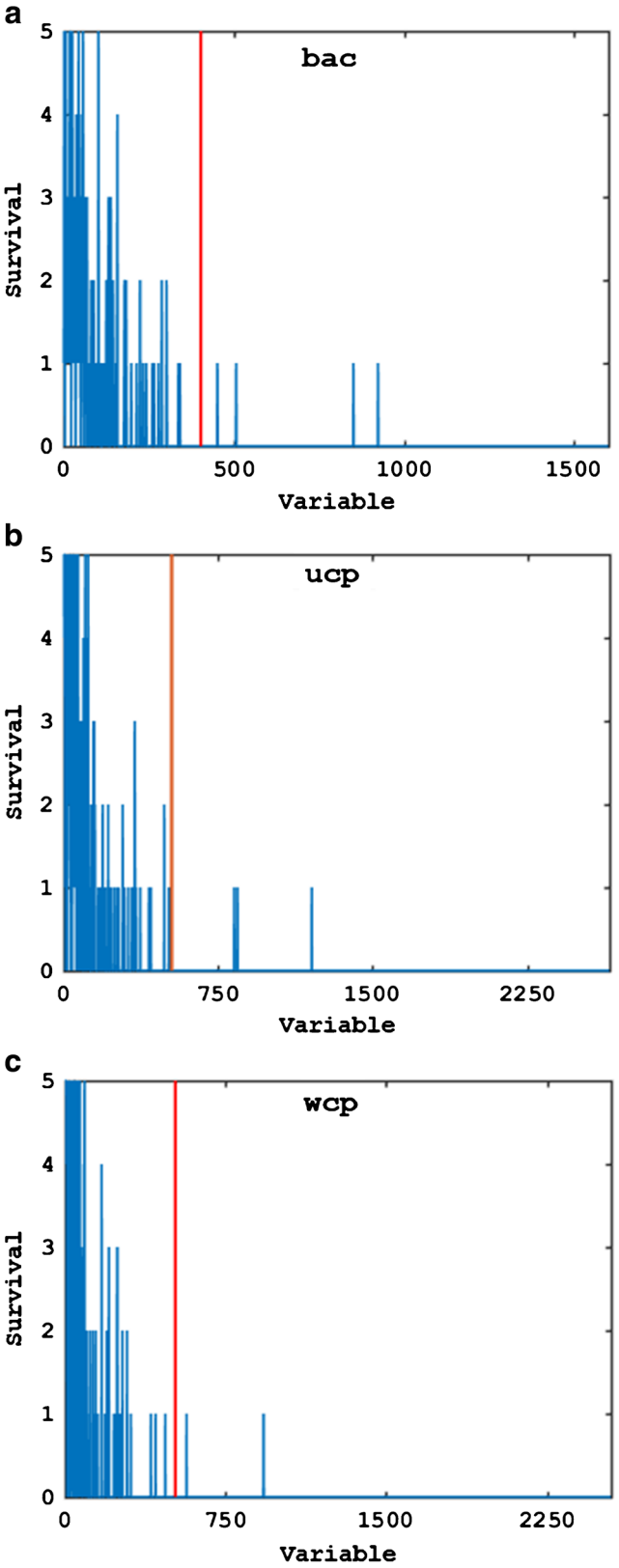

Fig. 5 Feature survival rate for the five values of d for **a** bac, **b** ucp, and **c** wcp. n_{SP} was estimated from the optimum d value; all variables were evaluated. Only features that survived at least 6 iterations are shown

Since n_{SP} was estimated from the empirical equation, it was also important to check if the values are below the optimum, i.e., was SFS being stopped too early. To check this, CR-FS was implemented using n_{ST} estimated for five d values from

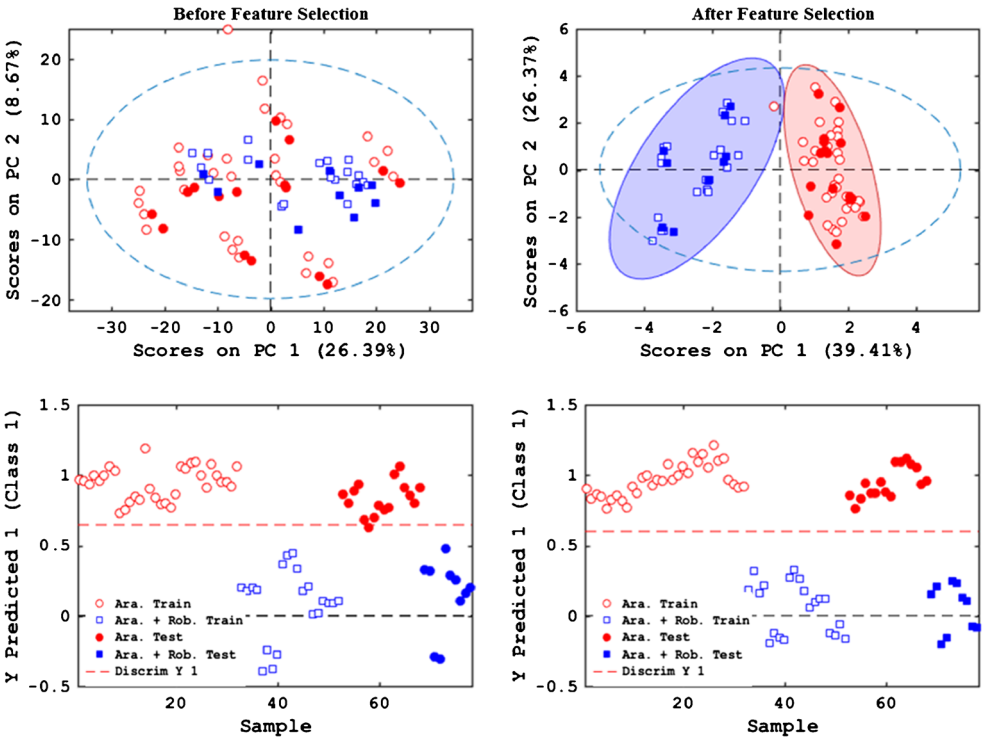

Fig. 6 PCA and PLS-DA models of coffee data to be classified as Arabica vs. a mixture (Arabica and Robusta). PCA and PLS-DA models before and after feature selection using 701 and 13 variables, respectively. Feature selection were performed using n_{ST} and n_{SP} estimated from Eqs. 2 and 3, respectively. Red markers represents Arabica, while blue markers represent mixture (Arabica and Robusta) coffee samples. Hollow and filled markers represents training and validation set data, respectively

0.48 to 0.57 with n_{SP} set to be equal to the total number of variables. For each of the five values of d, F-ratio analysis and the feature selection with CR-FS were performed ten times. During each of the ten iterations, a different subset of training data was selected for F-ratio analysis and model optimization. Feature survival rate was calculated as the number of times a variable is retained after CR-FS. A feature with a survival rate of 100% indicates that it was selected in all the ten iterations performed for each of the five d values, i.e., 50 times.

Figure 4a–c shows the overall feature survival rate bac, ucp, and wcp datasets. The n_{SP} estimated from F-ratio analysis of the data is indicated in red. A sharp drop in feature survival right after where the search should have been stopped (red line) can be seen.

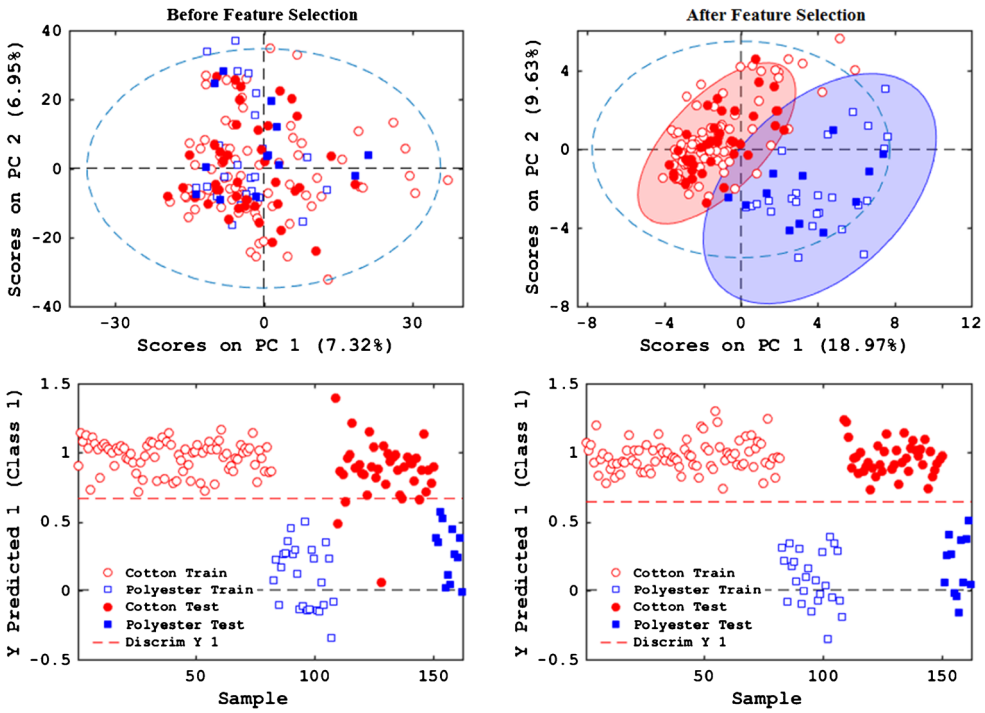

Fig. 7 PCA and PLS-DA models for fabric data to be classified as cotton vs. polyester. PCA and PLS-DA models before and after feature selection using 2781 and 35 variables, respectively. Feature selection were performed using n_{ST} and n_{SP} estimated from Eqs. 2 and 3, respectively. Red markers represents cotton while blue markers represent polyester samples. Hollow and filled markers represents training and validation set data, respectively

If each of the ten replicates for the values of d is treated as an independent feature selection, and a survival threshold of six, as used for earlier PLS-DA and PCA plots, then the feature survival plot is as shown in Fig. 5. In Fig. 5a–c, it can be seen that in all the three datasets (i.e., bac, ucp, and wcp), beyond n_{SP}, no variable survives at more than one d value. This indicates that the n_{SP} obtained from F-ratio analysis is a very good estimate.

Finally, CR-FS was implemented on two new datasets, 4 (cof) and 5 (cvp). Each of the datasets was split into 2/3 for training and 1/3 for validation sets. Using the training set, true and null ratio analysis was performed using half of the training set data. n_{ST} and n_{SP} were estimated from the F-ratio analysis using $d = 0.48$. The plot of PDFs for datasets 4 and 5 are shown in the ESM, Figs. S3 and S6. This was followed by feature selection with the CR-FS algorithm. The process was repeated ten times using a different subset of the training set data for the F-ratio analysis. For the coffee data, n_{ST} and n_{SP} were 17 and 160, respectively, while for the cvp data the values were 28 and 580, respectively. Features that survived at least six times were used to construct PCA and PLS-DA models. For the coffee data, 13 out of 160 features met this criteria. Figure 6 compares the PCA and PLS-DA results for coffee data before (106 features) and after feature selection (13 features). In the PCA model, the explained variance in the first and second principal components for before and after feature selection were from 35.06 and 75.78%, respectively. The prediction accuracy for the PLS-DA model improved from 96.3 to 100% after feature selection. For the cvp data, 35 out of 580 features survived. The PCA and PLS-DA models for the cvp data before (580 features) and after feature selection (35 features) are shown in Fig. 7. The PCA model shows an increase in the explained variance for the first two principal components from 14.27 to 28.60%. The PLS-DA model prediction accuracy for before and after feature selection was 90 and 100%, respectively.

Conclusions

Through the analysis of true and null F-ratios obtained from a dataset for classification models, an empirical equation was developed to estimate the start and stop numbers for CR-FS. All but one of the parameters in this equation are obtained by comparing the probability density functions of the true and null F-ratios. The constant to be determined was estimated to be in the range of $0.48 \leq d \leq 0.57$. The validity of this empirical equation was confirmed by testing on two new datasets. Using start and stop numbers obtained from the empirical equations, excellent model prediction accuracies were achieved with variables obtained after implementation of CR-FS. The use of this empirical equation can now be used as a guidance in setting the start and stop number for CR-FS, enabling a true automation of the feature selection process.

Acknowledgements The authors wish to acknowledge the Natural Sciences and Engineering Research Council of Canada (NSERC), Genome Canada and Genome Alberta, as well as Cystic Fibrosis Foundation Postdoctoral Fellowship (Bean12F0) and CF Isolate Core at Seattle Children's Research Institute (NIH P30 DK089507) for funding this research. They also wish to thank Dr. Aiko Barsch (Bruker Daltonics) for the coffee data used in this study.

Compliance with ethical standards Prior to any research being carried out involving human participants, all research protocols were approved by the relevant Human Research Ethics Board at the University of Alberta, including obtaining the informed consent of all participants in the wear trial that generated the fabric samples.

Conflict of interest The authors declare that they have no conflicts of interest.

References

1. Park J. Analogue and digital signals: practical data acquisition instrument control. System. 2003:13–35.
2. Measurement computing. Data acquisition handbook, a reference for DAQ and analog & digital signal conditioning. Third edit. A reference for DAQ And analog & digital signal conditioning. 2012.
3. Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big data: astronomical or genomical? PLoS Biol. 2015;13(7):1–11.
4. Wold S. Chemometrics; what do we mean with it, and what do we want from it? Chemom Intell Lab Syst. 1995;30(1):109–15.
5. Otto M. Chemometrics, statistics and computer application in analytical chemistry. 2nd ed. Weinheim: Wiley VCH; 2007.
6. Lavine BK. Source identification of underground fuel spills by pattern recognition analysis. Anal Chem. 1995;67(27):3846–52.
7. Malmquist LMV, Olsen RR, Hansen AB, Andersen O, Christensen JH. Assessment of oil weathering by gas chromatography-mass spectrometry, time warping and principal component analysis. J Chromatogr A. 2007;1164(1–2):262–70.
8. Nelson RK, Kile BM, Plata DL, Sylva SP, Xu L, Reddy CM, et al. Tracking the weathering of an oil spill with comprehensive two-dimensional gas chromatography. Environ Forensic. 2006;7(1):33–44.
9. Pasupuleti D, Eiceman GA, Pierce KM. Classification of biodiesel and fuel blends using gas chromatography—differential mobility spectrometry with cluster analysis and isolation of C18:3 me by dual ion filtering. Talanta. 2016;155:278–88.
10. Sigman ME, Williams MR, Castelbuono JA, Colca JG, Clark CD. Ignitable liquid classification and identification using the summed-ion mass spectrum. Instrum Sci Technol. 2008;36(4):375–93.
11. Sinkov NA, Sandercock PML, Harynuk JJ. Chemometric classification of casework arson samples based on gasoline content. Forensic Sci Int. 2014;235:24–31.
12. Lopatka M, Sampat AA, Jonkers S, Adutwum LA, Mol HGJ, van der Weg G, et al. Local ion signatures (LIS) for comparison of comprehensive two-dimensional gas chromatography applied to fire debris analysis. Forensic Chem. 2016;3:1–13.
13. Waddell EE, Song ET, Rinke CN, Williams MR, Sigman ME. Progress toward the determination of correct classification rates in fire debris analysis. J Forensic Sci. 2013;58(4):887–96.
14. Lopatka M, Sigman ME, Sjerps MJ, Williams MR, Vivo-Truyols G. Class-conditional feature modeling for ignitable liquid classification with substantial substrate contribution in fire debris analysis. Forensic Sci Int. 2015;252:177–86.

15. Farag MA, Otify A, Porzel A, Michel CG, Elsayed A, Wessjohann LA. Comparative metabolite profiling and fingerprinting of genus Passiflora leaves using a multiplex approach of UPLC-MS and NMR analyzed by chemometric tools. Anal Bioanal Chem. 2016;408(12):3125–43.
16. Xiao Z, Liu S, Gu Y, Xu N, Shang Y, Zhu J. Discrimination of cherry wines based on their sensory properties and aromatic fingerprinting using HS-SPME-GC-MS and multivariate analysis. J Food Sci. 2014;79(3):C284–94.
17. Cordero C, Kiefl J, Schieberle P, Reichenbach SE, Bicchi C. Comprehensive two-dimensional gas chromatography and food sensory properties: potential and challenges. Anal Bioanal Chem. 2014;407(1):169–91.
18. Debska B, Guzowska-Swider B. Decision trees in selection of featured determined food quality. Anal Chim Acta. 2011;705(1–2):261–71.
19. Guan W, Zhou M, Hampton CY, Benigno BB, Walker LD, Gray A, et al. Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. BMC Bioinf. 2009;10:259.
20. Szymanska E, Markuszewski MJ, Capron X, van Nederkassel AM, Vander Heyden Y, Markuszewski M, et al. Increasing conclusiveness of metabonomic studies by cheminformatic preprocessing of capillary electrophoretic data on urinary nucleoside profiles. J Pharm Biomed Anal. 2007;43(2):413–20.
21. Das MK, Bishwal SC, Das A, Dabral D, Varshney A, Badireddy VK, et al. Investigation of gender-specific exhaled breath volatome in humans by GCxGC-TOF-MS. Anal Chem. 2014;86(2):1229–37.
22. Katajamaa M, Orešič M. Data processing for mass spectrometry-based metabolomics. J Chromatogr A. 2007;1158(1–2):318–28.
23. Rajalahti T, Arneberg R, Berven FS, Myhr KM-M, Ulvik RJ, Kvalheim OM. Biomarker discovery in mass spectral profiles by means of selectivity ratio plot. Chemom Intell Lab Syst. 2009;95(1):35–48.
24. Shin H, Sheu B, Joseph M, Markey MK. Guilt-by-association feature selection: identifying biomarkers from proteomic profiles. J Biomed Inform. 2008;41(1):124–36.
25. Dang NA, Kolk AHJ, Kuijper S, Janssen H-G, Vivo-Truyols G. The identification of biomarkers differentiating Mycobacterium tuberculosis and non-tuberculous mycobacteria via thermally assisted hydrolysis and methylation gas chromatography-mass spectrometry and chemometrics. Metabolomics. 2013;9(6):1274–85.
26. Guyon I. An introduction to variable and feature selection 1 introduction. J Mach Learn Res. 2003;3:1157–82.
27. Guyon I, Elisseeff A. Feature extraction, foundations and applications: an introduction to feature extraction. Stud Fuzziness Soft Comput. 2006;207:1–25.
28. Engel J, Gerretzen J, Szymańska E, Jansen JJ, Downey G, Blanchet L, et al. Breaking with trends in pre-processing? TrAC Trends Anal Chem. 2013;50:96–106.
29. Bro R, Smilde AK. Centering and scaling in component analysis. J Chemom. 2003;17(1):16–33.
30. van den Berg RA, HCJ H, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006;7:142.
31. Craig A, Cloarec O, Holmes E, Nicholson JK, Lindon JC. Scaling and normalization effects in NMR spectroscopic metabonomic data sets. Anal Chem. 2006;78(7):2262–7.
32. Caruana RA, Freitag D. How useful is relevance? AAAI Fall Syposium on Relevance. New Orleans; 1994. 25–9.
33. John GH, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. 11th International Conference on Machine Learning. New Brunswick; 1994. 121–9.
34. John GH, Kohavi R. Wrappers for feature subset selection. Artif Intell. 1997;97(1):273–324.
35. Hall M. Correlation-based feature selection for machine learning. Methodology. 1999:1–5.
36. Vieira SM, Sousa JMCC, Kaymak U. Fuzzy criteria for feature selection. Fuzzy Sets Syst. 2012;189(1):1–18.
37. Boser BE, Guyon IM, Vapnik VN. Training algorithm for optimal margin classifiers. Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory; 1992. 144–52.
38. Sánchez-Maroño N, Alonso-Betanzos A, Tombilla-Snaromán M. Filter methods for feature selection—a comparative study. Intell Data Eng Autom Learn – IDEAL. 2007;178–87.
39. Science C, Arabia S. Learning boolean concepts in the presence of many irrelevant features. Artif Intell. 1994;69:279–305.
40. Cadenas JM, Garrido MC, Martínez R. Feature subset selection filter–wrapper based on low quality data. Expert Syst Appl. 2013;40(16):6241–52.
41. Soufan O, Kleftogiannis D, Kalnis P, Bajic VB. DWFS: a wrapper feature selection tool based on a parallel genetic algorithm. PLoS One. 2015;10(2):1–23. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0117988
42. Rinke CN, Williams MR, Brown C, Baudelet M, Richardson M, Sigman ME. Discriminant analysis in the presence of interferences: combined application of target factor analysis and a Bayesian soft-classifier. Anal Chim Acta, Elsevier BV. 2012;753:19–26.
43. Farrés M, Platikanov S, Tsakovski S, Tauler R. Comparison of the variable importance in projection (VIP) and of the selectivity ratio (SR) methods for variable selection and interpretation. J Chemom [Internet]. 2015;29(10):528–36. Available from: http://doi.wiley.com/10.1002/cem.2736
44. Rajalahti T, Arneberg R, Kroksveen AC, Berle M, Myhr K-M-M, Kvalheim OM. Discriminating variable test and selectivity ratio plot: quantitative tools for interpretation and variable (biomarker) selection in complex spectral or chromatographic profiles. Anal Chem. 2009;81(7):2581–90.
45. Sinkov NA, Harynuk JJ. Cluster resolution: a metric for automated, objective and optimized feature selection in chemometric modeling. Talanta [Internet], Elsevier B.V. 2011;83(4):1079–87.
46. Sinkov NA, Harynuk JJ. Three-dimensional cluster resolution for guiding automatic chemometric model optimization. Talanta. 2013;103:252–9.
47. Johnson KJ, Synovec RE. Pattern recognition of jet fuels: comprehensive GC×GC with ANOVA-based feature selection and principal component analysis. Chemom Intell Lab Syst. 2002;60(1–2):225–37.
48. Adutwum LAA, Harynuk JJJ. Unique ion filter: a data reduction tool for GC/MS data preprocessing prior to chemometric analysis. Anal Chem Am Chem Soc. 2014;86(15):7726–33.
49. de la Mata AP, McQueen RH, Nam SL, Harynuk JJ. Comprehensive two-dimensional gas chromatographic profiling and chemometric interpretation of the volatile profiles of sweat in knit fabrics. Anal Bioanal Chem. 2017;409(7):1905–13.
50. Oliynyk AOO, Adutwum LAA, Harynuk JJJ, Mar A. Classifying crystal structures of binary compounds AB through cluster resolution feature selection and support vector machine analysis. Chem Mater. 2016;28(18):6672–81.
51. Parsons BA, Marney LC, Siegler WC, Hoggard JC, Wright BW, Synovec RE. Tile-based Fisher ratio analysis of comprehensive two-dimensional gas chromatography time-of-flight mass spectrometry (GC×GC-TOFMS) data using a null distribution approach. Anal Chem. 2015;87(7):3812–9.
52. Weitzman MS. Measures of overlap of income distributions of white and Negro families in the United States. US Bureau of the Census; 1970.
53. Inman HF, Bradley EL. The overlapping coefficient as a measure of agreement between probability distributions and point estimation of

the overlap of two normal densities. Commun Stat Theory Methods. 1989;18(10):3851–74.
54. Matusita K. Decision rule, based on the distance, for the classification problem. Ann Inst Stat Math. 1956;8(1):67.
55. Mulekar MS, Mishra SN. Confidence interval estimation of overlap: equal means case. Comput Stat Data Anal. 2000;34(2):121–37.
56. Akaike H. Information theory and an extensión of the maximum likelihood principle. Int Symp Inf Theory. 1973;1973:267–81.
57. Hu S. Akaike information criterion statistics. Math Comput Simul. 1987;29(5):452.
58. Tellstrom V, Harder A, Barsch A. Metabolic profiling of different coffee types on the Bruker compactTM QTOF system. Application Note. 2013. Available from: https://www.bruker.com/fileadmin/user_upload/8-PDF-Docs/Separations_MassSpectrometry/Literature/literature/ApplicationNotes/LCMS-79_compact_QTOF_03-2013_eBook.pdf
59. DeLeeuw J. Introduction to Akaike (1973) information theory and an extension of the maximum likelihood principle. Breakthroughs in statistics volume I: foundations and basic theory. 1992. p. 599–609.
60. Snipes M, Taylor DC. Model selection and Akaike information criteria: an example from wine ratings and prices. Wine Econ Policy. 2014;3(1):3–9.