Talanta 103 (2013) 252–259

Contents lists available at SciVerse ScienceDirect

Talanta

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

Three-dimensional cluster resolution for guiding automatic chemometric model optimization

Nikolai A. Sinkov, James J. Harynuk*

Department of Chemistry, University of Alberta, Edmonton, AB, Canada T6G TG2

ARTICLE INFO

Article history:
Received 7 July 2012
Received in revised form
11 October 2012
Accepted 11 October 2012
Available online 26 October 2012

Keywords:
Chemometrics
Cluster resolution
PCA
GC–MS
Gasoline
Selectivity ratio
Variable selection

ABSTRACT

A three-dimensional extension of a previously developed metric termed *cluster resolution* is presented. The cluster resolution metric considers confidence ellipses (here three-dimensional confidence ellipsoids) around clusters of points in principal component or latent variable space. Cluster resolution is defined as the maximum confidence limit at which confidence ellipses do not overlap and can serve to guide automated variable selection processes. Previously, this metric has been used to guide variable selection in a two-dimensional projection of data. In this study, the metric is refined to simultaneously consider the shapes of clusters of points in a three-dimensional space. We couple it with selectivity ratio-based variable ranking and a combined backward elimination/forward selection strategy to demonstrate its use for the automated optimization of a six-class PCA model of gasoline by vendor and octane rating. Within-class variability was artificially increased through evaporative weathering and intentional contamination of samples, making the optimization more challenging. Our approach was successful in identifying a small subset of variables (644) from the raw GC–MS chromatographic data which comprised $\sim 2 \times 10^6$ variables per sample. In the final model there was clear separation between all classes. Computational time for this completely automated variable selection was 36 h; slower than solving the same problem using three two-dimensional projections, but yielding an overall better model. By simultaneously considering three dimensions instead of only two at a time, the resulting overall cluster resolution was improved.

1. Introduction

Chemometric techniques are invaluable tools for the interpretation of complex analytical data. For example, chemometrics are used to determine the origin of samples or identify changes in samples over time as a result of exposure to certain conditions [1,2]. Chemometric techniques are used in a diversity of fields. In food science, chemometric techniques have been applied to the analysis of olive oils [3,4], the determination of fatty acids in cow's milk using FT-IR [5] and to chromatographic analysis of natural products [6]. Chemometrics can also be used in forensics, for example to aid in fingerprinting ignitable fluids and determining their origins during arson investigations [7–9]. In the field of metabolomics, chemometrics have been applied to a variety of chromatography–mass spectrometry data, including recent applications to GC–MS [10], LC-MS [11,12] and UPLC-MS data [13,14].

Chromatography, especially when hyphenated to mass spectrometry, provides incredibly rich data. Chemometric techniques can use this volume of data to their advantage, and are becoming more popular options for data interpretation. However, the richness of the data is also their downfall. Prior to chemometric analysis, some pre-processing that incorporates a data reduction step must be applied to minimize the number of irrelevant variables subsequently input into chemometric tools. One option is to integrate the signal and perform chemometric analysis on the integrated peak table [1,2,7–9,15,16]. An advantage of this approach is that data matrix obtained is relatively simple. However, this approach is only viable in instances where there are no coelutions or where analytes can be deconvoluted based on their mass spectra. For complex samples, this may be impossible.

Alternatively, one can directly apply chemometric techniques to raw chromatographic signals [3–6,11,13,17–26]. The challenge with this approach (apart from data alignment) is that most variables will only include random noise, especially when a MS is used as the detector. Consequently, a strategy is required to select a subset of variables from the original data that are likely to contain the most relevant information while ignoring as many irrelevant variables as possible.

Variable selection techniques for GC–MS can be relatively fast and computationally simple. Total ion current [18], or extracted ion chromatograms/profiles [15–18] can be used, though these strategies are likely to keep many uninformative variables in the

Abbreviations: CR, Cluster resolution; FS, Forward selection; BE, Backwards elimination; SR, Selectivity ratio; PCA, Principal component analysis; PC, Principal component; LV, Latent variable

* Corresponding author. Tel.: +1 780 492 8303; fax: +1 780 492 8231.
E-mail address: James.Harynuk@ualberta.ca (J.J. Harynuk).

0039-9140/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2012.10.040

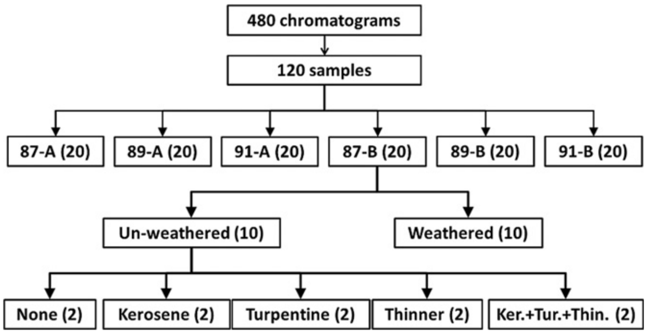

Fig. 1. Schematic for sample preparation.

model, while risking the loss of useful variables and the multivariate advantage of GC–MS data.

Hypothetically, an exhaustive search could be used; the strategy tests all possible combinations of variables and the best combination is chosen. While this method is guaranteed to find the best possible set of variables, it is only applicable to situations with a very small number of variables (i.e., integrated peak tables with few peaks). The number of possible combinations of variables is 2^n where n is the number of variables; testing all combinations of just 20 variables, assuming a calculation time of 100 ms for each iteration would require ~29 h. Obviously, a more efficient approach that targets a select population of variables is required, especially for the handling of raw chromatographic data.

One solution to this challenge is to rank variables based on their perceived utility in a future model and then choose a small group of top-ranked variables. Examples of such approaches include analysis of variance (ANOVA) [20,21,24,25] and selectivity ratio (SR) plots [6,27–29]. Using only some top-ranked variables, an improved model can be generated by avoiding those variables that include more noise than information. However each ranking metric may provide different results and no ranking metric is perfect. It is possible for some lower-ranked variables to contain useful information while some higher-ranked ones are harmful or redundant. In these cases, a subset of variables should be selected by some other means, though variable ranking may still play a role in reducing the population of considered variables to a manageable size.

Stepwise methods such as forward selection (FS) and backward elimination (BE) [30–32] are two possible methods for automated feature selection. In FS, variables are sequentially added to the model; if the addition of a variable is deemed to improve model quality, the variable is permanently retained. The process continues until a predefined endpoint is reached, such as checking a certain number of variables, including a certain number of variables, or reaching a certain value of model quality (according to some metric). In BE, the process starts with a model containing all possible candidate variables. Variables are then removed sequentially. If the removal of a variable does not degrade model quality, it is permanently excluded. The process continues until all variables have been checked. While BE is somewhat more computationally intensive than FS, its major advantage is that the influence of each variable is considered within the context of other variables [30].

Model evaluation is also a crucial step in optimization. Approaches to model evaluation include receiver operating characteristic (ROC) curves [33], or metrics based on Mahalanobis [19,34] and Euclidian distances [24,25], for example. In our previous works we have introduced and demonstrated the use of cluster resolution (CR) as a metric to quantify separation between classes in a model [20,35]. Advantages of CR include its ability to account for sizes, orientations, and positions of clusters of points in Principal Component (PC) or Latent Variable (LV) spaces. CR works by determining the maximum sized confidence ellipses that can be described about a pair of classes without overlap. In previous works, CR was applied to 2 PC and 2 LV models. Since chemometric analyses often involve working in a higher dimensional space, the two-dimensional limitation is addressed here. In this work, CR is expanded to consider clusters of points projected into a three-dimensional space.

2. Experimental

Gasoline samples were obtained from two local gas stations in Edmonton, Alberta, Canada. Each station belonged to a different vendor and three different octane ratings of gasoline (87, 89 and 91) were obtained from each vendor, providing a total of six classes. To introduce some challenge to the variable selection process, datasets were made more complicated by introducing a higher degree of within-class variance. To this end, half of the samples in each class were weathered approximately 50% by volume using a gentle stream of clean, dry, compressed air. To introduce further in-class variance, some samples from each class were left uncontaminated, some were contaminated by adding either turpentine (5% by volume), lacquer thinner (5% by volume), kerosene (5% by volume), or a mixture of turpentine, lacquer thinner and kerosene together (5% by volume each). A total of 120 samples were prepared and their compositions are shown in Fig. 1. The samples were then diluted 20:1 by volume in pentane and analyzed by GC–MS. The GC–MS used for these experiments was a 7890A GC with a 5975 quadrupole MS (Agilent Technologies, Mississauga, ON) equipped with a 30 m × 250 μm; 0.25 μm HP-5 column (Agilent). The carrier gas used was helium at a constant flow rate of 1.0 mL min^{-1}. The injector was held constant at 250 °C and a volume of 0.2 μL was injected using a split ratio of 100:1. The temperature program was 50 °C (3.5 min hold) with a 20 °C min^{-1} ramp to 300 °C. The total run time was 16 min. The initial solvent delay was 2.5 min and mass spectra were collected from m/z 30 to m/z 300 at the rate of 9.2 spectra s^{-1}.

Four chromatograms were collected from each sample, providing a total of 480 chromatograms (80 for each class). Chromatograms were then assigned to training, optimization and validation sets. From each class, 40 chromatograms were assigned to the training set, 20 were assigned to the optimization set and 20 were assigned to the validation set, to the total of 240 chromatograms in the training set, 120 chromatograms in optimization set and 120 chromatograms in the validation set. Chromatograms from the training set were used to create the alignment target, rank variables and to obtain loading vectors for the PCA model in each variable selection step. Chromatograms from the optimization set were used, together with chromatograms from the training set, to obtain scores during variable selection as well as to create the final PCA model after variable selection was complete. Chromatograms from the validation set were not used until after variable selection was finished and the final model was constructed, serving only to validate the final model.

For each analysis, the entire chromatogram was exported as a .csv file, which was then imported into MATLAB 7.10.0.499 (The Mathworks, Natick, MA) as a 7300 × 271 (scan number × m/z ratio) matrix using a lab-written algorithm. Data were then handled using lab-written algorithms. Chemometric models were constructed using PLS toolbox 5.8 (Eigenvector Research Inc., Wenatchee, WA). Chromatographic alignment was performed based on the piecewise alignment algorithm developed by Johnson et al. [36] with an additional mass spectral confirmation to match features, though in principle any alignment algorithm could be used. First, an alignment target was created.

The preliminary target was constructed by randomly choosing a chromatogram from the training set. Then, a second, randomly chosen chromatogram from the training set was aligned with the target chromatogram after which it was added to the preliminary target. Then, the aligned chromatogram was discarded and the algorithm proceeds to the next chromatogram in the training set. After all chromatograms from the training set had been included in the target, the algorithm proceeded with the alignment of all chromatograms in all sets (including the training set) to the composite target chromatogram.

3. Theory

In this study, we seek to optimize a PCA model as an example. PCA projects data onto a set of orthogonal vectors called principal components (PCs), reducing dimensionality of the data, which allows for its easier visualization and interpretation. The general equation for PCA is given by

$$X = \mathrm{TP} + E \tag{1}$$

where $\mathbf{X}$ is the original data matrix ($n \times m$), $\mathbf{T}$ is the scores matrix ($n \times k$), $\mathbf{P}$ is the loadings matrix ($k \times m$) and $\mathbf{E}$ is the residual matrix ($n \times m$). The number of samples in the dataset is represented by n, while m is the number of variables included for each sample, and k is the number of PCs used to construct the model.

While PCA by itself is not a classification technique [37], it can be used to project high-dimensional (say, $m = 1000$) data onto a relatively smaller number of PCs, allowing easier visualization of the dataset [38]. Unlike PLS-DA, PCA does not by itself provide false positive or false negative rates, posing a problem for optimizations relying on ROCs. CR has been successfully applied to two-dimensional PCA [20] as well as PLS-DA [35] models. It has also been performed with both ANOVA and SR-based variable ranking.

When treating raw GC–MS data, the recorded intensity for each ion in each scan is considered a separate variable. For example, each chromatogram in this study contained about two million individual variables, most of which contained no relevant information. To perform variable selection within a reasonable amount of time, the total number of variables considered must be decreased to a few thousand from a few million. Using a variable ranking technique, the variables most likely to be the most informative are identified, and a subset of the highest-ranked variables are chosen as candidates for inclusion.

In this study, SR was used as a variable ranking technique. Briefly, SR involves the creation of a PLS-DA model. Scores and loadings of the target-projected model are calculated and then the ratio of explained variance versus residual variance is calculated for each variable, providing the SR for that variable [27,28,39].

We have found CR to be an efficient guide for variable selection when the initial model is somewhat stable and has at least some separation between classes. Thus, in this study a combined BE/FS approach was applied as outlined in Fig. 2.

3.1. Cluster resolution

When PCA is performed on a dataset containing two or more different classes of samples, each class will ideally cluster in a different region of the scores plot. The size of each cluster will depend on the degree of variation within each class and the distance between each cluster will depend on how well the included features can describe the differences between the classes. Generally, it is desirable to have a model where classes are as far apart as possible on the scores plot, while samples within each class cluster together as tightly as possible. Important variables will contribute more towards increasing the distance between clusters while causing a minimal increase in the size of each cluster. Conversely, irrelevant variables will do little to increase the separation, but will render each cluster of samples more diffuse. CR was developed as a metric that accounts for the distance between clusters of points, while considering their relative orientations and sizes. Similarly to the Degree of Class Separation (DCS) metric [24,25], CR will measure how well clusters are separated relative to their sizes and orientations. The major difference is that while DCS describes clusters as spheres or circles, CR describes classes as ellipses or ellipsoids thus somewhat accounting for the shapes and relative orientations of the clusters. This has some advantages, especially as clusters of points describing each class often form shapes that are more elliptical than circular. The metric is based upon determination of the maximum confidence limit at which confidence ellipses described around clusters of points in PC or LV space are separated, and its application in two dimensions has been described previously [20,35].

3.2. Cluster resolution in three dimensions

Confidence ellipsoids can theoretically be created in any number of dimensions by constructing an n-component PCA model around a cluster of points defining the directions of the n axes of the ellipsoid. The axes can be combined with the confidence limit in each dimension to provide the size of the ellipsoid. With the position of the ellipsoid center, as well as sizes and directions of the ellipsoid axes, a confidence ellipsoid with approximately evenly-spaced points covering its surface is constructed. In this work we demonstrate three-dimensional ellipsoids ($n = 3$). With two such ellipsoids constructed around two clusters of samples, collision detection is performed. If a collision is detected, the confidence limit is reduced for the following iteration of collision detection. If a collision is not detected, the confidence limit is increased in the following iteration. The highest confidence limit at which confidence ellipses are still separated defines the CR. In a multi-class model, the calculation must be performed for each possible pairing of classes, and the product of CRs for all possible pairs of classes is the overall quality metric for the model.

4. Results and discussion

Presented here is a further development of the cluster resolution metric, the primary use of the metric is the evaluation of chemometric models during optimization, including but not limited to variable selection. Here, CR was used to guide a combined BE/FS variable selection process with the goal of constructing a three-component PCA model with the greatest degree of separation between clusters for each class. In our example, the data consist of 80 GC–MS chromatograms for each of six types of gasoline (three octane ratings from each of two vendors), to the total of 480 chromatograms. In order to increase the challenge for the algorithm, within-class variability was increased by weathering and/or contaminating some gasoline samples (Fig. 1). The 480 chromatograms from each class were randomly split into a training set (40 chromatograms from each class), a validation set (20 chromatograms from each class) and a test set (20 chromatograms from each class), after which chromatographic alignment was performed.

The aligned matrices were then unfolded along the time axis to yield a series of vectors. Each vector consisted of $\sim 2 \times 10^6$ variables. SR variable ranking was applied to the set of 240 chromatograms in the training set using a lab-written algorithm.

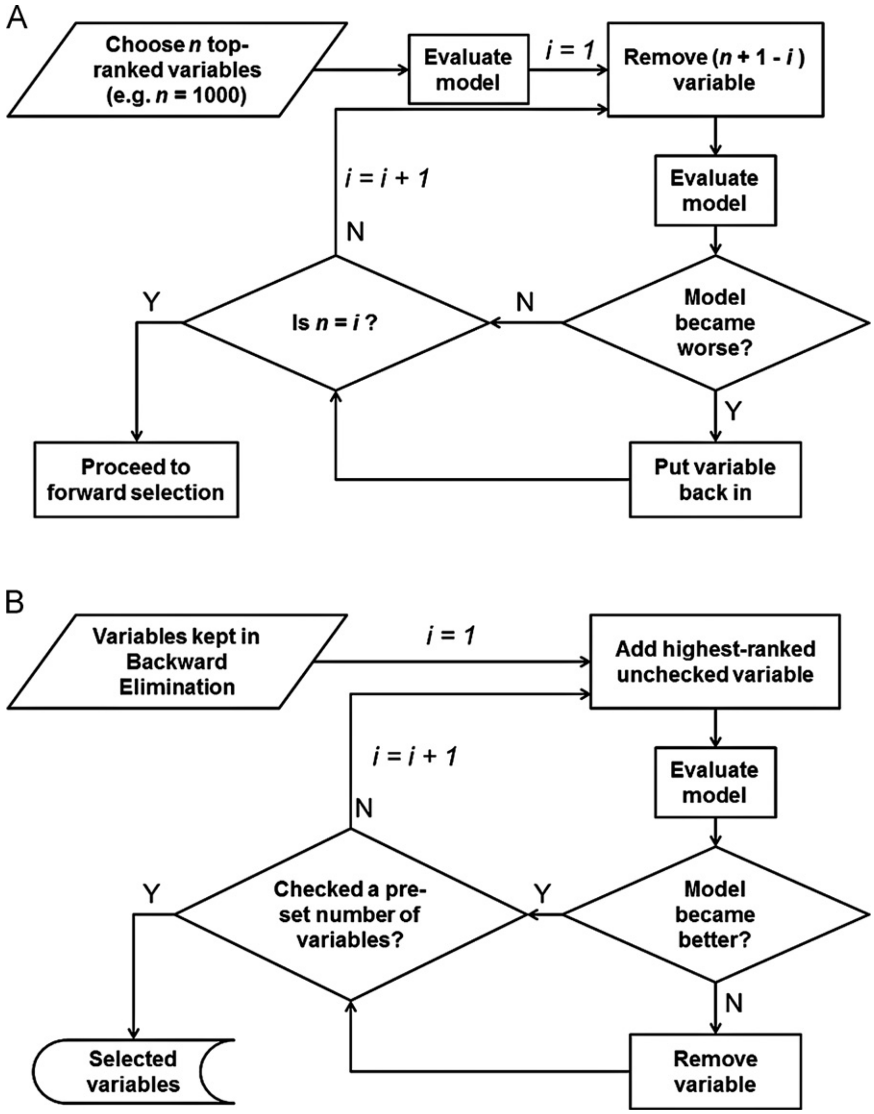
A
Choose n top-ranked variables (e.g. n = 1000)
Evaluate model
i = 1
Remove (n + 1 - i) variable
Evaluate model
i = i + 1
N
Y
Is n = i ?
N
Model became worse?
Y
Proceed to forward selection
Put variable back in
B
Variables kept in Backward Elimination
i = 1
Add highest-ranked unchecked variable
Evaluate model
i = i + 1
N
Y
Checked a pre-set number of variables?
Y
Model became better?
N
Selected variables
Remove variable

Fig. 2. Variable selection techniques used. (A) Backwards elimination; (B) Forward selection. CR used in the *Evaluate Model* step.

This yielded a vector of selectivity ratios that was used to rank the features. The optimization and test data-sets were aligned as well, but were not used in the calculation of selectivity ratios. Baseline correction was not necessary as the variable ranking process automatically down-weights background ions which did not vary significantly from sample to sample. Computation of the SR ranking vector from the aligned data required approximately one minute.

After variable ranking, the 1000 top-ranked variables were selected and a three-component PCA model was created using the training set. Using the scores from both the training and optimization sets on the first 3 PCs, six clusters of points with 60 points per cluster were obtained. Overall model quality was calculated as the product of the individually determined CR measurements for each of the 15 possible of pairings of clusters. BE was performed on the 1000 top-ranked variables (Fig. 2A). The variables retained after BE were then passed to FS where variables ranked 1001 through 3000 were considered for inclusion (Fig. 2B). Variable selection took approximately 36 h to complete and selected a total of 644 variables from the 3000 variables checked.

The training and optimization sets (360 chromatograms) were then combined to train the final three-component PCA model using normalization to an area of 1 and autoscaling as the only pre-processing methods. Subsequent projection of the validation set (120 as yet unused chromatograms) permitted evaluation of the final model. Fig. 3A depicts the resulting model where lightly shaded regions are three-dimensional 98% confidence ellipsoids described around clusters of training set samples (individual points not shown) and markers represent individual points for validation set samples. Colours represent different classes. As can be seen from the figure, validation set samples projected into the same regions as training set samples and all classes were separated in the three-dimensional space. Overall, the final measured three-dimensional CR for this problem was 0.9997.

The three-dimensional CR metric was compared with the previously developed two-dimensional CR metric. The same training, optimization and validation sets were used and, just as in the three-dimensional case, backwards elimination started with the 1000 top-ranked variables (as shown in Fig. 2A) and forward selection checked variables ranked 1001 through 3000 (as shown in Fig. 2B). To make the comparison fair, a three-component PCA model was constructed at each step and the two-dimensional CR metric was calculated for the three possible two-dimensional projections of the three principal components. For each pair of classes, the CR value retained to guide optimization was the highest value among the three projections (e.g., for

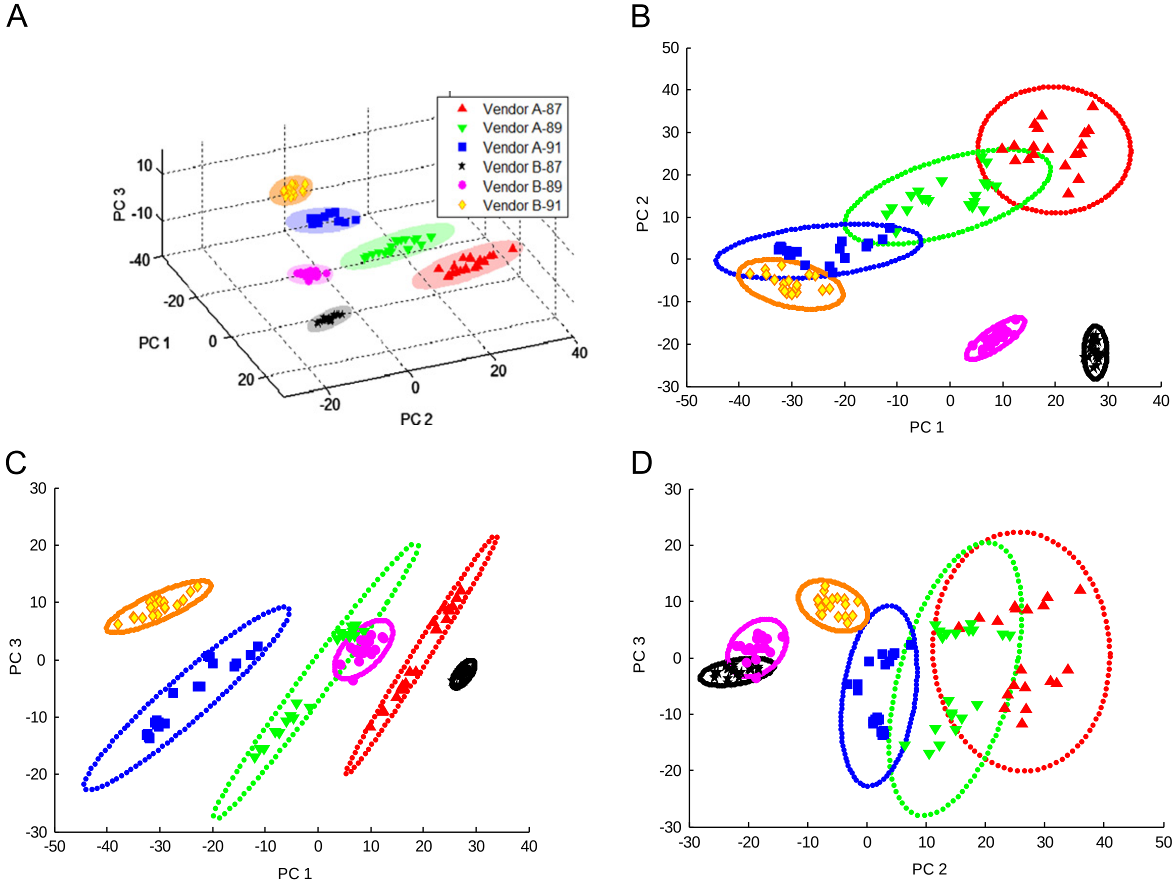

Fig. 3. Scores plots for the final PCA model. Red, green, and blue represent 87-, 89-, and 91-octane gasoline from Vendor A, respectively. Black, magenta, and orange represent 87-, 89-, and 91-octane gasoline from Vendor B, respectively. (A) shaded regions represent three-dimensional confidence ellipsoids (98%) described around clusters of training set samples (individual points not shown), solid markers represent individual points for validation set samples. (B)–(D) dotted lines represent two-dimensional confidence ellipses at 98% confidence limit described around clusters of training set samples (individual points not shown) and markers represent individual points for validation set samples on projection of three-dimensional model onto components 1 and 2, 1 and 3, 2 and 3, respectively. (For interpretation of the references to color in this figure legend, the reader is refferred to the web version of this article.)

discriminating the 91-octane gasolines from Vendor A and Vendor B (blue and orange classes) in Fig. 3 B–D, the CR score on PC1 vs. PC3 was retained). This calculation required about 12 h. Fig. 3B–D represent two-dimensional projections of the optimized model showing 98% confidence ellipses based on training set samples (individual points not shown) and individual markers for validation set samples. As can be seen, no individual two-dimensional projection was able to separate all classes of all samples though using the three projections together, separation was achieved using a total of 1009 variables. The final two-dimensional CR, calculated based on the best projection for each pairing, was 0.9985. When the three-dimensional CR was calculated for a model based on the selected 1009 variables, it was found to be 0.9991. Thus, the three-dimensional CR approach yielded a better model than the two-dimensional CR approach in this case (though the practical difference between CR values of 0.9991 and 0.9997 is worthy of future study).

Once variables have been selected, they can be traced back to the original data and tentative identities of the selected compounds can be postulated. Here, we used mass spectral information combined with linear retention indices [40] for compound identification. Fig. 4 depicts a binary mask where black dots represent the 644 variables selected when using the three-dimensional CR-guided approach and white space represents excluded variables. As a comparison, three 89-octane samples are presented in Figs. 5 and 6, with Fig. 5 depicting a region of the raw GC–MS chromatograms and Fig. 6 depicting the abundances of the selected variables in each chromatogram (Fig. 5 masked by Fig. 4).

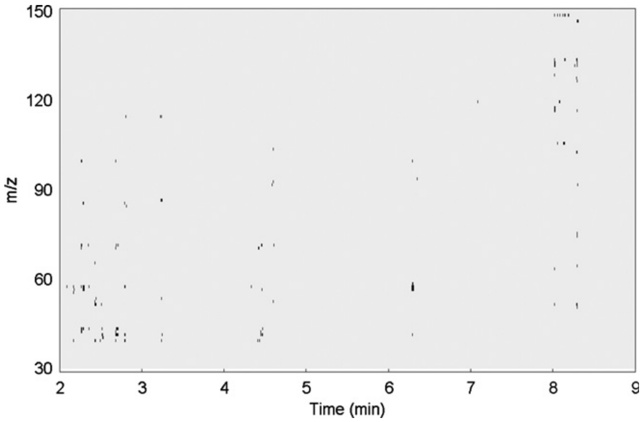

Fig. 4. Variables selected from the original data by the algorithm. Black dots represent variables that were selected after BE/FS.

The gasoline samples from the two vendors can be distinguished on the basis of several compounds. First of all, Vendor A (Fig. 6A) has a relatively high abundance of C5 alkylbenzenes (eluting between 8 and 9 min) whereas Vendor B (Fig. 6B and C) does not have an appreciable amount of these compounds.

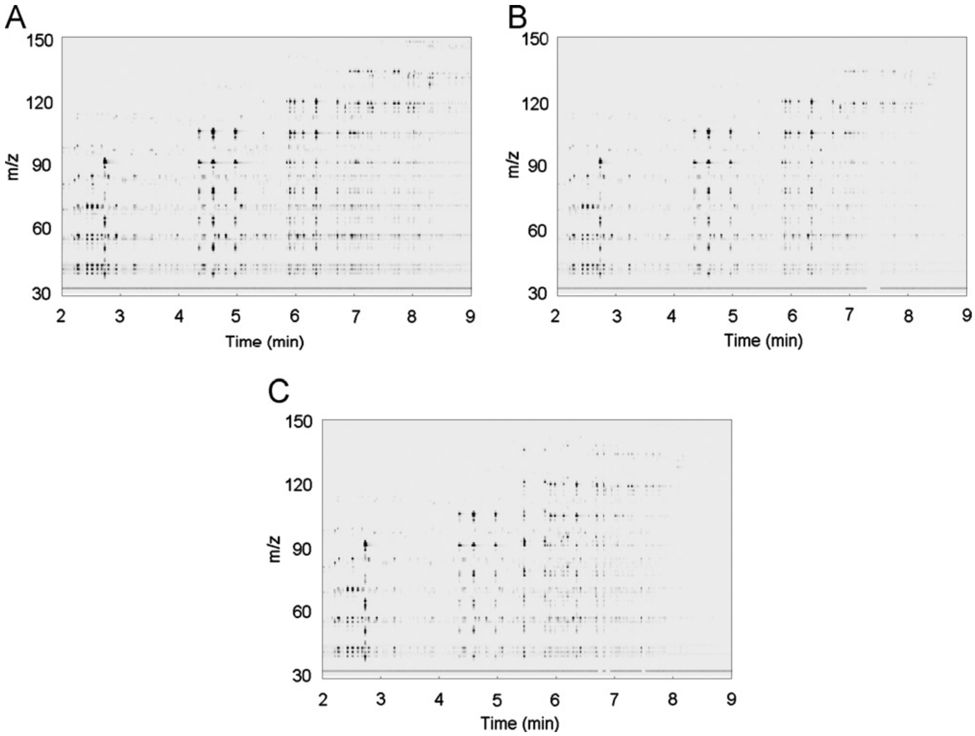

Fig. 5. GC–MS chromatograms of selected gasoline samples after alignment was performed. Light grey indicates low signal while dark grey indicates high signal for a variable. (A) Vendor A 89-octane weathered uncontaminated gasoline. (B) Vendor B 89-octane weathered uncontaminated gasoline. (C) Vendor B 89-octane unweathered gasoline contaminated with kerosene, turpentine and lacquer thinner.

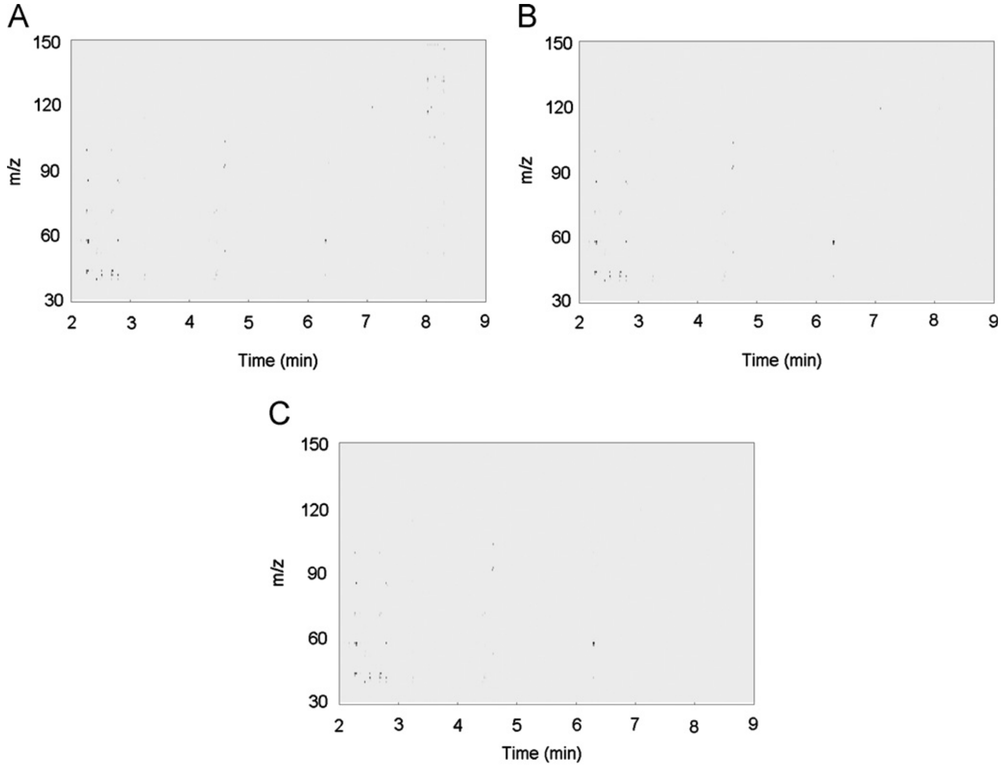

Fig. 6. Variables selected for GC–MS chromatograms of selected gasoline samples after alignment was performed. Light grey indicates low signal while dark grey indicates high signal for a variable. Variables that were not selected were assigned value of zero. (A) Vendor A 89-octane weathered uncontaminated gasoline. (B) Vendor B 89-octane weathered uncontaminated gasoline. (C) Vendor B 89-octane unweathered gasoline contaminated with kerosene, turpentine and lacquer thinner.

Additionally, Vendor A has a slightly increased abundance of a peak at 4.4 min and a much lower abundance of a compound at 6.3 min. These two compounds have been tentatively identified as 4-methyl octane and 2,2,4,6,6-pentamethylheptane on the basis of their mass spectral matches and their linear retention indices (Table 1).

Table 1
Tentative identification of two compounds relevant for distinguishing between classes of samples. Calculated linear retention index was obtained from the data, literature linear retention index was obtained from reference [40], forward and reverse mass spectral matches were obtained by comparing the background-corrected peak apex mass spectrum with the NIST database spectrum for a given compound.

Compound	Experimental LRI	Literature LRI	MS match (Forward)	MS match (Reverse)
4-methyl octane	868	864	850	889
2,2,4,6,6-pentamethylheptane	992	997	876	879

Considering two samples from the same class but with added within-class variance, we see that there is a significant difference between unweathered contaminated Vendor B 89-octane, and weathered, uncontaminated gasoline from the same class (Fig. 5B and C, respectively). However, when one considers only the selected variables, the two samples (which belong to the same class: Vendor B, 89-octane) are essentially identical (Fig. 6B and C). Upon closer inspection of Fig. 4, it can be seen that between 2 and 3 min, signals due to several alkanes were selected. This is consistent with the fact that light alkanes are present at very different levels in gasolines of different types. Toluene is commonly present in gasoline and, differences in toluene abundance have been previously shown to be useful in discriminating between different classes of gasoline [20]. Under our conditions, toluene is found to elute at approximately 2.7 min, and is shown to be an an abundant compound in gasoline. As seen in Figs. 4 and 6, toluene is not selected as a useful variable. This is explained by the high concentration of toluene in the lacquer thinner that was added as a contaminant. Thus, in this data set, toluene contributes significantly to within-class variation, decreasing its utility for distinguishing the classes of gasoline. The feature selection algorithm automatically discovered this and correctly discarded toluene from the model.

A final point for this discussion is computation time. It took approximately 36 h to perform variable selection, with cluster resolution being the slowest step. However, it should be noted that with n classes there are $(n^2/2)-n$ possible pairs of classes; here 15 pairs were considered. Thus it only required slightly more than 2 h of computation time for each pair of classes. Since the calculation of CR for each pair of classes is independent from the calculation for each other pair, this step could be easily distributed and calculated in parallel across multiple processors, greatly speeding up computation time. It should also be noted that variable selection was completely automated. Once the classes were assigned to the data files, the remainder of the process concluded with no user intervention or attention required.

5. Conclusions

Three-dimensional CR is a further development of the CR metric. It has been shown to serve as an effective guide for automated variable selection and model optimization. For this particular data set, both two- and three-dimensional CR metrics worked to guide feature selection to similar optimal models. The two-dimensional approach was faster to compute, while the three-dimensional approach yielded a slightly better model. Even though this data set could be optimized quite effectively by both approaches, conceptually, situations are imaginable where the three-dimensional approach would yield a significantly better model than the two-dimensional approach. This would be due to the simultaneous consideration of three-dimensional ellipsoids permitting the optimization of more complex systems where using one or more two-dimensional projections would likely fail. While in this study CR was applied in conjunction with FS and BE, both two- and three-dimensional CR measurements can be easily used as “goodness metrics” in other variable selection approaches.

Acknowledgements

The authors would like to acknowledge the Department of Chemistry at the University of Alberta, and the Natural Sciences and Engineering Research Council (NSERC) Canada for their support of this research. Alberta Innovates Technology Futures' support through a New Faculty Award to Dr. Harynuk is also acknowledged.

References

[1] B.M. Zorzetti, J.M. Shaver, J.J. Harynuk, Anal. Chim. Acta 694 (2008) 31–37.
[2] P. Doble, M. Sandercock, E. Du Pasquier, P. Petocz, C. Roux, M. Dawson, Forensic Sci. Int. 132 (2003) 26–39.
[3] P. de la Mata-Espinosa, J.M. Bosque-Sendra, R. Bro, L. Cuadros-Rodriguez, Talanta 85 (2011) 177–182.
[4] P. de la Mata, A. Dominguez-Vidal, J.M. Bosque-Sendra, A. Ruiz-Medina, L. Cuadros-Rodríguez, M.J. Ayora-Cañada, Food Control 23 (2012) 449–455.
[5] M. Ferrand, B. Huquet, S. Barbey, S. Barillet, F. Faucon, H. Larroque, O. Leray, J.M. Trommenschlager, M. Brochard, Chemom. Intell. Lab. Syst. 106 (2011) 183–189.
[6] O.M. Kvalheim, H.Y. Chan, I.F.F. Benzie, Y.T. Szeto, A.H.C. Tzang, D.K.W. Mok, F.T. Chau, Chemom. Intell. Lab. Syst. 107 (2011) 98–105.
[7] P.M.L. Sandercock, E. Du Pasquier, Forensic Sci. Int. 134 (2003) 1–10.
[8] P.M.L. Sandercock, E. Du Pasquier, Forensic Sci. Int. 140 (2004) 43–59.
[9] P.M.L. Sandercock, E. Du Pasquier, Forensic Sci. Int. 140 (2004) 71–77.
[10] I. Oliver, D.T. Loots, J. Microbiol. Methods 88 (2012) 419–426.
[11] I.D. Wilson, R. Plumb, J. Granger, H. Major, R. Williams, E.M. Lenz, J. Chromatogr. B 817 (2004) 67–76.
[12] A. Kiss, A.L. Jacquet, O. Paisse, M.M. Flament-Waton, J. de Ceaurriz, C. Bordes, J.Y. Gauvrit, P. Lanteri, C. Cren-Olive, Talanta 83 (2011) 1769–1773.
[13] S.J. Bruce, P. Johnsson, H. Antti, O. Cloarec, J. Trygg, S.L. Marklund, T. Moritz, Anal. Biochem. 372 (2008) 237–249.
[14] Y. Hou, D.R. Braun, C.R. Michel, J.L. Klassen, N. Adnani, T.P. Wyche, T.S. Bugni, Anal. Chem. 84 (2012) 4277–4283.
[15] B.T. Weldegergis, A.M. Crouch, J. Agric. Food Chem. 56 (2008) 10225–10236.
[16] R.B. Gaines, G.J. Hall, G.S. Frysinger, W.R. Gronlund, K.L. Juaire, Environ. Forensics 7 (2006) 77–87.
[17] C.R. Borges, Anal. Chem. 79 (2007) 4805–4813.
[18] L.J. Marshall, J.W. McIlroy, V.L. McGuffin, S.R. Waddell, Anal. Bioanal. Chem. 394 (2009) 2049–2059.
[19] J.H. Christensen, G.J. Tomasi, Chromatogr. A 1169 (2007) 1–22.
[20] N.A. Sinkov, J.J. Harynuk, Talanta 83 (2011) 1079–1087.
[21] N.E. Watson, M.M. VanWingerden, K.M. Pierce, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1129 (2006) 111–118.
[22] R.E. Mohler, B.P. Tu, K.M. Dombeck, J.C. Hoggard, E.T. Young, R.E. Synovec, J. Chromatogr. A 1186 (2008) 401–411.
[23] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Analyst 132 (2007) 756–767.
[24] K.J. Johnson, R.E. Synovec, Chemom. Intell. Lab. Syst. 60 (2002) 225–237.
[25] K.M. Pierce, J.L. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1096 (2005) 101–110.
[26] J.S. Nadeau, B.W. Wright, R.E. Synovec, Talanta 81 (2010) 120–128.
[27] T. Rajalahti, R. Arneberg, F.S. Berven, K.M. Myhr, R.J. Ulvik, O.M. Kvalheim, Chemom. Intell. Lab. Syst. 95 (2009) 35–48.
[28] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, Anal. Chem. 81 (2009) 2581–2590.
[29] J.P.M. Andries, Y. Vander Heyden, L.M.C. Buydens, Anal. Chim. Acta 705 (2011) 292–305.
[30] I. Guyon, A.J. Elisseeff, Mac. Learn. Res. 3 (2003) 1157–1182.
[31] R.R. Hocking, Biometrics 32 (1976) 1–49.
[32] D.E. Axelson, Data Preprocessing for Chemometric and Metabonomic Analysis, first ed., MRi Consulting, Kingston, Ontario, 2010.

[33] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Metabolomics 4 (2008) 81–89.

[34] J.H. Christensen, A.B. Hansen, U. Karlson, J. Mortensen, O. Andersen, J. Chromatogr. A 1090 (2005) 133–145.

[35] N.A. Sinkov, B.M. Johnston, P.M.L. Sandercock, J.J. Harynuk, Anal. Chim. Acta 697 (2011) 8–15.

[36] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J Chromatogr. A 996 (2003) 141–155.

[37] J.M. Bosque-Sendra, L. Cuadros-Rodriguez, C. Ruiz-Samblas, A.P. de la Mata, Anal. Chim. Acta 724 (2012) 1–11.

[38] S. Wold, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.

[39] O.M. Kvalheim, Chemom. Intell. Lab. Syst. 8 (1990) 59–67.

[40] X. Xu, L.L.P. van Stee, J. Williams, J. Beens, M. Adahchour, R.J.J. Vreuls, U.A.T. Brinkman, J. Lelieveld, Atmos. Chem. Phys. 3 (2003) 665–682.