Talanta 83 (2011) 1079–1087

Contents lists available at ScienceDirect

Talanta

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

# Cluster resolution: A metric for automated, objective and optimized feature selection in chemometric modeling

Nikolai A. Sinkov, James J. Harynuk*

*Department of Chemistry, University of Alberta, Edmonton, Alberta, T6G 2G2, Canada*

ARTICLE INFO

*Article history:*
Available online 27 October 2010

*Keywords:*
Cluster resolution
Feature selection
Chemometrics
PCA
ANOVA
GC–MS
Gasoline

ABSTRACT

A novel metric termed cluster resolution is presented. This metric compares the separation of clusters of data points while simultaneously considering the shapes of the clusters and their relative orientations. Using cluster resolution in conjunction with an objective variable ranking metric allows for fully automated feature selection for the construction of chemometric models. The metric is based upon considering the maximum size of confidence ellipses around clusters of points representing different classes of objects that can be constructed without any overlap of the ellipses. For demonstration purposes we utilized PCA to classify samples of gasoline based upon their octane rating. The entire GC–MS chromatogram of each sample comprising over $2 \times 10^6$ variables was considered. As an example, automated ranking by ANOVA was applied followed by a forward selection approach to choose variables for inclusion. This approach can be generally applied to feature selection for a variety of applications and represents a significant step towards the development of fully automated, objective construction of chemometric models.



## 1. Introduction

There are many well-established chemometric techniques used to facilitate the handling of chemical data: techniques such as principal components analysis (PCA) and partial least-squares (PLS) being among the most common. This proliferation of chemometric techniques can be attributed to several factors, including improvements in computing technology and more user-friendly software coupled with advancements in analytical instrumentation. Modern instrumental techniques, especially the families of hyphenated separations techniques (e.g.: GC–MS and LC–MS) and multidimensional separations, tend to be data-rich and provide an abundance of detail pertaining to the nature of a complex sample in a relatively short period of time. This in turn has permitted the analyst to probe increasingly complex samples, and pose increasingly challenging questions, the answers to which are most easily revealed though the use of chemometric tools. For example, grades of gasoline were classified based on their GC–MS profiles by Doble et al. using PCA [1]. Sandercock and Du Pasquer have used GC–MS coupled with PCA to fingerprint a series of gasoline samples and identify the origin of the samples [2–4].

Another field where chemometric techniques are widely applied is in metabolomics (as well as general metabolite profiling and metabonomics). Wilson et al. have recently reviewed the application of LC–MS in this field, highlighting some uses of chemometrics [5]. Other examples of the use of chemometrics in this area include Bruce et al. who recently evaluated metabolite profiling techniques and used PCA and orthogonal projections to latent structures discriminant analysis (OPLS-DA) on UPLC–MS data [6]. Chemometric techniques have also been used in conjunction with metabolic data for the prediction of gender [7], the early detection of cancer [8], classification of tobacco extracts [9], and the study of yeast metabolites [10,11].

When applying chemometric techniques to chromatographic or chromatographic–mass spectrometric data, there are several possible approaches to preparing the data for analysis. Many users employ integrated peak tables of data as this provides a matrix that is relatively small and straightforward: analyte abundances vs. sample numbers [1–4,7,12–14]. Other users choose to use a non-integrated chromatographic signal for the construction of a chemometric model [5,6,8,10,11,15–26]. With this approach, each variable in the data matrix is the signal intensity at a given time. This route has its own challenges, including increased data size and data alignment; however, these can be overcome relatively easily and this approach is in many cases superior to the use of integrated peak tables. The advantage of using the entire raw data set is more evident when one utilizes the entire GC–MS chromatogram, either as a three-way array (scan number $\times$ $m/z$ ratio $\times$ sample number) or as a two-dimensional array of samples vs. GC–MS chromatograms unfolded along their time axis. Synovec and co-workers

* 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 © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.talanta.2010.10.025

demonstrated the significant advantages can be achieved by using the entire GC–MS chromatogram rather than extracted ion chromatograms or other univariate signals [22]. The reason for this being that the chemometric model can extract underlying patterns in the data that are not evident in univariate signals.

One challenge that remains for all types of chemometric analyses is that of feature selection: choosing which of the variables that have been collected will be included in the chemometric model. In cases where one is utilizing raw chromatographic data or chromatographic–mass spectrometric data, feature selection becomes at the same time more challenging and more important as millions of variables can be easily collected for each sample. A dataset comprising even a relatively small number of samples such as these will put inordinate demands on a computer system. Apart from the technological challenge, the most important reason for careful variable selection is that not all variables will be relevant. This is especially true when the entire chromatogram is considered: only a small portion of the chromatographic space actually contains relevant signal intensities. If irrelevant variables are included, the model must account for irrelevant variations and this will degrade its overall performance. Consequently, careful variable selection is necessary, especially if raw chromatographic signals are being used in the construction of the model [18,27].

There are multiple variable selection techniques that are available, all with the goal of simplifying data sets and removing extraneous variables. Selection techniques such as using integrated peak tables [13,14], extracted ion chromatograms [15,16] (EIC), or single ion monitoring [19] (SIM) rely very highly on *a priori* knowledge and select variables by only permitting a small, user-selected portion of the data to be used. These are not inappropriate approaches, but they are potentially dangerous, especially if the system is poorly understood. The reason being that in these cases there are numerous opportunities for either the inclusion of significant quantities of irrelevant data or the inadvertent exclusion of relevant portions of the data. Within the scope of chromatography–MS data, total ion chromatograms (TICs) may also be used for modeling [16,20], but this sacrifices essentially all of the additional mass spectral information and potentially useful variables in the process.

Objective variable ranking techniques are another option for guiding feature selection. These methods use a calculated metric to evaluate the potential value of each variable. When constructing a model, only those variables with scores above a certain threshold will be used. Examples of the use of objective variable ranking applied to chromatographic data include the work of Rajalahti et al. where the discriminating variable (DIVA) test was used to rank variables for both PCA and PLS-DA of chromatographic profiles [21]. Another popular metric for variable ranking is Analysis of Variance (ANOVA) which has been used to guide feature selection for PCA of GC–MS and GC × GC chromatograms [18,22,23]. Teófilo et al. have also used informative vectors as the ranking metric prior to PLS analysis of spectroscopic data [24]. Apart from the inherent advantage of objectivity, objective ranking strategies allow the user to consider many more candidate variables with no *a priori* information and can be readily incorporated into automated routines.

Another approach to variable selection is the application of a genetic algorithm (GA). The main advantage of GAs is that they can proceed without much user intervention. However, they are computationally expensive, typically exhibit severe overfitting of the data and/or converge to non-optimal solutions, especially with data sets that comprise a large number of variables. Strategies to overcome these limitations have recently been presented by Ballabo et al. [17]. However, GAs remain comparatively computationally inefficient. Additional feature selection approaches exist, but within the scope of gas chromatography the above methods are the most common.

It should also be noted that the challenge of feature selection is by no means limited to the field of chromatography, or even chemistry. For example, uncorrelated linear discriminant analysis (ULDA)-based feature selection has been applied to both classification of cancer samples and biomarker discovery using time-of-flight mass spectrometry (TOFMS) data [28], and in the field of economics, multivariate discriminant analysis (MDA) was employed to identify predictors of business failure [29]. Regardless of the variable selection technique that is applied, the goals are to remove noise and irrelevant variables while preserving variables that are of value. For example, when techniques such as ANOVA and DIVA are used to select variables to be included in a PCA model, variables are ranked based on their relative ability to discriminate between the classes of samples being considered. Variables with a high ranking are likely to improve class separation, and those with a low ranking are deemed to be irrelevant. As more variables are included, it is more likely that information useful for class discrimination will be included in the model, though each additional variable is likely to be less useful than the previous ones [27]. However, with each new variable more noise is added to the model, possibly reducing the model's ability to discriminate between classes. At some point, the addition of new variables will result in an overall loss of model quality.

This highlights the central problem that we address in this research. In cases where one is attempting to construct a chemometric model of a large data set, how one objectively choose an optimal combination of features to model the data? Further, how can one quantify and thereby objectively compare the separation and clustering of data points belonging to multiple classes in, for example, a PCA model? This can be judged through visual inspection of various diagnostic plots of the model; however, in order to achieve a fully automated and objective process for feature selection, an objective metric is required.

In this study we present such a metric. While metrics for the degree of class separation have been used previously [19,23,25], prior metrics do not account simultaneously for the shapes, sizes and relative orientations of clusters of points on, for example, a PCA scores plot. The metric that we have developed has been termed *cluster resolution* and it considers these three parameters, representing a significant advancement over previous metrics. We also compare the use of this metric and a metric based on Euclidean distances in an algorithm to automatically construct a PCA model for classifying a series of gasoline samples based upon their GC–MS profiles. It must be noted that this metric may be used to compare any two models where the separation between groups of clusters needs to be evaluated. It can equally be applied to partial least squares discriminant analysis or factor analysis, for example. Additionally, the way in which variables are added is also flexible. Here we demonstrate the approach with forward selection based on ANOVA rankings as this is computationally trivial and straightforward, but informative vectors, DIVA, genetic algorithms, or any other selection method could equally be used.

## 2. Experimental

To demonstrate cluster resolution and its use in automated feature selection, a test set of gasoline samples was used. Three gasoline samples having octane ratings of 87, 89, and 91 were obtained from a single local gas station in Edmonton, Alberta, Canada. The samples were diluted 20:1 (v/v) 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 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 with 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.

A total of 24 chromatograms were collected for each of the gasoline samples over a period of 2 weeks. The entire mass chromatogram for each analysis was exported as a .csv file, which was then imported into MATLAB 7.10.0 (The Mathworks, Natick, MA) as a 7400 × 271 (scan number × $m/z$ ratio) matrix using a lab-written algorithm. Data were then handled in MATLAB using lab-written algorithms. Chemometric models were constructed using PLS toolbox 5.2 (Eigenvector Research Inc., Wenatchee, WA).

## 3. Theory

Cluster resolution is a metric that evaluates the distance between clusters of points, relative to their sizes. The metric may be applied with any one of many combinations of feature selection and modeling approaches. In discussing the development of the cluster resolution metric, we will use the example of PCA as this is one of the most widely used techniques for data visualization and exploration. Briefly, PCA projects multivariate data with a high dimensionality into a series of orthogonal subspaces, allowing for quick and easy visualization and interpretation of highly dimensional data in a lower-dimensional subspace [30,31]. The first principal component (PC) explains the most variance within the dataset, and each subsequent component describes less and less variance.

While the number of PCs to include in the model is chosen by the user, in practice data are often viewed as plots of scores on combinations of two or three PC axes. When PCA is performed on a dataset containing 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 within-class variation and the distance between each cluster will depend on how well the variability in the included features can describe the differences between the classes. In general 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 as tightly together as possible. As stated previously, the inclusion of additional relevant variables will drive clusters farther apart, while the addition of less important variables will do little to increase the separation, but will render each cluster of samples more diffuse. In order to automate the variable selection process, a calculable metric must be available that can account for the distance between clusters, while considering their relative orientations and sizes.

Metrics for the degree of between-class separation exist in the literature. One straightforward manner in which to measure the separation between classes is to compare the Euclidean distance between the centroids of a pair of classes, relative to square root of the sum of the variance within each group [18,23]. Another metric compares the sum of Mahalanobis distances between samples belonging to each class and the centroid of the class with the sum of Mahalanobis distances between all samples and the model origin [19,32]. While these metrics do permit an estimate of between-class separation relative to the sizes of the clusters, they suffer from the fact that they do not consider the shapes and orientations of each class. As demonstrated by Fig. 1, the relative orientation of a pair of ellipses can have a critical impact on whether they are separated or not at a given confidence limit. While the clusters in Fig. 1A are clearly separated in case of 75, 95 and 99% confidence limits, in Fig. 1B only the 75% confidence limits are separated, even though

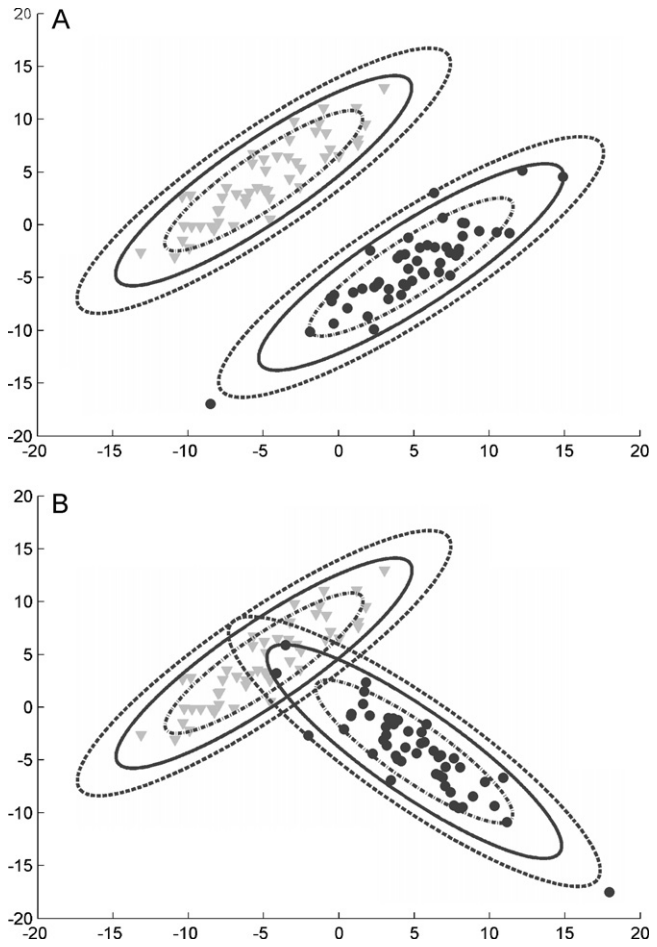


**Fig. 1.** Two clusters of points with 75, 95 and 99% confidence ellipses. (A) Ellipses are oriented parallel to each other, and (B) ellipses oriented such that they have some overlap. The centroids of the ellipses and the sizes of the ellipses have not changed.

the centroids of these two clusters as well as variance within each class were identical in both cases. In fact, to the metric based on Euclidean distance measurements, these two cases have an identical test statistic. To the cluster resolution metric, the cases in Fig. 1 are easily distinguishable, demonstrating a more accurate measure of class separation and a superior metric for automating the process of variable selection.

Another advantage of cluster resolution over metrics such as Euclidean or Mahalanobis distances is that cluster resolution is bounded between 0 and 1. Consequently it is very easy to evaluate the overall model quality when more than one class of samples exists. This is achieved by simply taking the product of cluster resolutions for all pairings of ellipses. Additional terms in the product of ellipses which are very well separated (cluster resolution >0.95) will leave the overall model quality high, but even a single poor resolution (e.g.: 0.45) will dominate the product term, indicating overall poor model quality. This will be demonstrated with our experimental data.

### 3.1. Calculation of cluster resolution

Cluster resolution is defined as the maximum confidence limit at which confidence ellipses describing two different classes are still separated. In cases where more than two classes exist, cluster resolution is calculated for each pair of ellipses, and then the product of these values is used to indicate overall model performance. An algorithm has been developed in our group to reliably and automatically determine that limit.

Once a PCA model has been constructed, scores for each class on a pair of PC axes are used for evaluating the models; here we will use PC1 and PC2. Considering only the data from a single class, the scores for the points in the cluster provide new variables from which a 2-component PCA model is constructed without any data scaling. The resultant loading vectors for PCs 1 and 2 provide the directions of the major and minor axes for the resulting confidence ellipse that describes this class. The eigenvalues, along with critical Hotelling $T^2$ value for a given confidence limit, the number of samples in the class, and number of PCs describing the cluster (here two as we are using a two-dimensional confidence ellipse), provide the lengths of each of the ellipse's axes. This process is repeated for each cluster of points in the model.

To construct a confidence ellipse for a class, the covariance matrix of scores is first calculated:

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}'\mathbf{X} \tag{1}$$

where **S** is the scores covariance matrix, $n$ is the number of samples in the cluster being evaluated, and **X** is a matrix where the rows are the samples in a cluster and the columns are the scores of each sample on PCs 1 and 2.

Singular Value Decomposition (SVD) is then performed on the covariance matrix:

$$\mathbf{S} = \mathbf{ULV}' \tag{2}$$

where **U** and **V**′ are identical and provide loading vectors for the model describing the confidence ellipse, while $L$ is a diagonal matrix containing eigenvalues for components 1 and 2 of the new model. The number of samples within the cluster permits the determination of the Hotelling $T^2$ value for a given confidence limit [33]:

$$T^2 = \frac{p(n-1)}{n-p}F(\alpha, p, n-p) \tag{3}$$

where $p$ is number of components in a model (in this case 2), $n$ is the number of samples in the class, $\alpha$ is the confidence limit and $F(\alpha,p,n-p)$ is the $F$ statistic for given values of $\alpha$, $p$ and $n$. The length of each confidence ellipse axis ($l$) is given by Eq. (4).

$$l = \sqrt{T^2 \times L} \tag{4}$$

When $L$ is the eigenvalue for PC 1 the length of the major axis is provided, and when $L$ is the eigenvalue of PC 2, the length of the minor axis is provided. With both directions and lengths of the axes describing ellipses at a given confidence limit calculated for each cluster of points, a set of approximately 1000 evenly spaced points are distributed along the circumference of each ellipse. This is achieved by warping a circle comprising 100 000 points until it is superimposable on the confidence ellipse. To reduce the number of points in the ellipse, its circumference is calculated using Rajmanujan's approximation [34] and divided by 1000 to yield the distance between two adjacent points, $d$. Then, beginning at an arbitrary point on the circumference of the ellipse, the algorithm proceeds along the ellipse until a point a distance $d$ along the ellipse is found. Points between the starting point and this second point are discarded. This process is repeated around the entire ellipse with the result being an ellipse with about 1000 points distributed evenly along its circumference. The choice of 1000 points was made because it provides a balance between accurate representation of the ellipses and computational speed/requirements.

To determine if two confidence ellipses overlap at a given confidence level, the Euclidean distances between all points on one ellipse and those on a second ellipse are calculated. The minimum of these distances ($D_{min}$) is compared to half the sum of the distances between two neighbouring points on the circumferences of each ellipse ($D_{critical}$). If the minimum distance, $D_{min}$, is less than the critical distance, $D_{critical}$, the two ellipses are deemed to overlap.

To determine the maximum confidence limit at which ellipses will not overlap, the algorithm begins with an arbitrary confidence limit (in this work we chose to use 75%) and determines if there is any overlap. If overlap is detected, the algorithm decreases the confidence limit for both ellipses in a stepwise fashion until overlap is no longer detected. Conversely, if there is no overlap detected, the algorithm increases the confidence limits of the two ellipses until overlap is detected. The highest confidence limit at which there is no overlap detected is defined as the cluster resolution for a given pair of classes. Cluster resolution is calculated for each pair of classes separately and has values above 0 and below 1 (representing 0 and 100% confidence ellipses).

During feature selection, variables are added to the model in a forward-selection process. The first time that cluster resolution is determined; the algorithm begins from the arbitrary confidence limit (75%). However, after the first iteration, the cluster resolution for each pair of clusters is stored and used as the starting point for subsequent iterations with additional variables.

### 3.2. Application of cluster resolution in automated variable selection

The initial step in automated variable selection is to rank the variables according to some metric, for example ANOVA or DIVA. The choice of ranking metric for the purpose of demonstrating the application of cluster resolution is arbitrary, though it should be noted that in our work we have observed that different ranking metrics produce different models exhibiting maximal cluster resolution. A comparison of ranking methods is beyond the scope of the present discussion, but is a topic for future study. Here we used ANOVA, which has been described and demonstrated previously [18,22,23,26], as it is computationally inexpensive and straightforward. The two main limitations in using ANOVA are that it assumes that the observed variance is normally distributed, and that when used on a data set where the number of variables vastly exceeds the number of samples (such as will be the case whenever raw chromatographic data are subjected to ANOVA) it is entirely possible for ANOVA to find features that can discriminate between classes based upon random fluctuations in the data as opposed to meaningful variances. A thorough discussion of these limitations is beyond the scope of this paper, though the easiest manner for one guard against these problems is to use a large training set and to perform a cross-validation of the results with another data set.

Briefly, the output of ANOVA is a series of $F$ ratios for each variable. The $F$ ratio is a measure of the ratio of between-class variance to within-class variance. If a variable has an elevated $F$ ratio, then it is deemed to be more valuable for describing the difference between classes. With the $F$ ratio calculated for every data point in the chromatogram, the variables are ranked in order of decreasing $F$ ratio. A PCA model is then constructed using a fraction of variables that have the highest $F$ ratio. In principle any number of PCs could be used, though for the sake of computational speed, it is best to use as few components as possible. For this proof-of-concept work, a two-component model was used.

Once a model is constructed, each class of sample will occupy a given region on the scores plot, and a confidence ellipse can be described around each cluster. The cluster resolution between each pair of classes is then calculated. This process is repeated, including more and more variables until the desired endpoint is reached, using the previously determined cluster resolution for each pair of classes as the starting point for evaluating the cluster resolution with additional variables. The endpoint may be the number of variables where the resolution is maximized (such as when the critical pair of classes shows the highest cluster resolution or when product of all cluster resolutions for all class pairs is maximized), or it may be when the minimum resolution is greater than a threshold value,

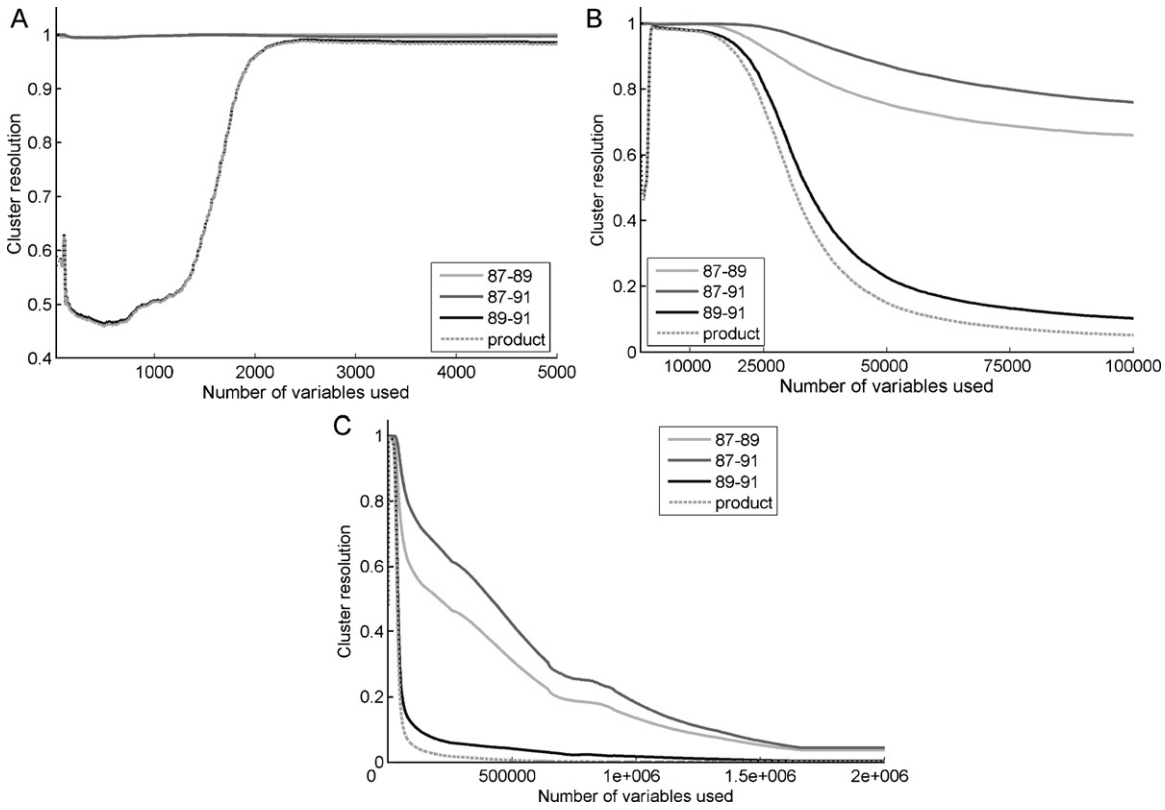


**Fig. 2.** Resolution of gasoline clusters as a function of the number of variables used for Set 1. (A) Close up of the region from 0 to 5000 included variables; (B) close up of the region from 0 to 100 000 included variables; (C) full resolution plot from 0 to 2 000 000 included variables.

for example 0.95 (meaning that no confidence ellipses exhibit overlap at the 95% confidence level). Herein we look to maximize the poorest resolution and product of resolutions.

## 4. Results and discussion

Presented here is the proof-of-concept for a novel metric that we term cluster resolution. While the purpose of the metric is the objective evaluation of the separation between clusters of points, it can be used to automate the feature selection process and compare similar chemometric models in an objective manner. As a demonstration, we have used cluster resolution in an algorithm to automatically select the features in the data which can be used to construct the PCA model having the greatest degree of separation between clusters for each class. In our example, the data comprise 72 GC–MS chromatograms of gasolines having three different octane ratings. The 72 chromatograms were randomly split into a training set (containing 16 chromatograms from each class) and a test set (containing the remaining 8 chromatograms from each class). This was repeated four times to use a total of five different randomly chosen training and test sets to evaluate the stability of the solution to minor variations in the training data. Finally the procedure was performed on the complete set of data with no test set.

For data alignment, chromatograms were aligned using a homemade alignment function based upon the piecewise alignment algorithm developed by Synovec and co-workers [35], with an additional mass spectral confirmation of features to be matched, though in principle any alignment algorithm could be used. The target to which data were aligned was a composite chromatogram of a series of aligned gasoline samples of different octane ratings. This ensured that all components present in the samples were present in the alignment target, though not necessarily at the same abundances. The aligned matrices were then unfolded along the time axis to yield a series of vectors. ANOVA was applied to the set of 48 chromatograms in the each training set using a lab-written algorithm. For each set, this yielded a vector of $F$ ratios that was used to rank the features. The test data sets were aligned as well, but were not used in calculation of $F$ ratios. Baseline correction was not necessary as the ANOVA process automatically down-weights background ions which do not vary significantly from sample to sample.

With variables now ranked by their $F$ ratios, the data in the training sets were autoscaled and subsets of data containing all rows (samples) and the desired number of columns (features) were extracted and used to construct a two-component PCA model. The cluster resolution between each possible pairing of classes on the scores plot for PC1 vs. PC2 was then calculated on the basis of the training data set. This step was repeated sequentially, adding more and more variables at each step to find the optimal number of variables to include in the PCA model for each training set.

The original training data set comprised a matrix with 48 rows (representing samples), and 2 005 400 columns (representing variables). In each case, the maximum number of variables to be included was limited to 100 000. In one case calculations were performed to include up to the entire set of $2 \times 10^6$ variables to demonstrate the problem with utilizing the entire raw data file, especially when it is incredibly sparse (Fig. 2C). In terms of computational time, it may take a few minutes to calculate the initial cluster resolution, depending on the data and the step size used for changing the confidence limits as the algorithm must "walk" from the arbitrary value. However, as the resolutions that are found in an iteration are used as the starting points for the subsequent iteration, the speed is limited by how fast variables can be extracted from a dataset and the PCA model is constructed. In practice this is about two seconds per step. In order to efficiently determine the optimal number of variables to use, a large step size can be used in the first pass through the data to find the approximate location of

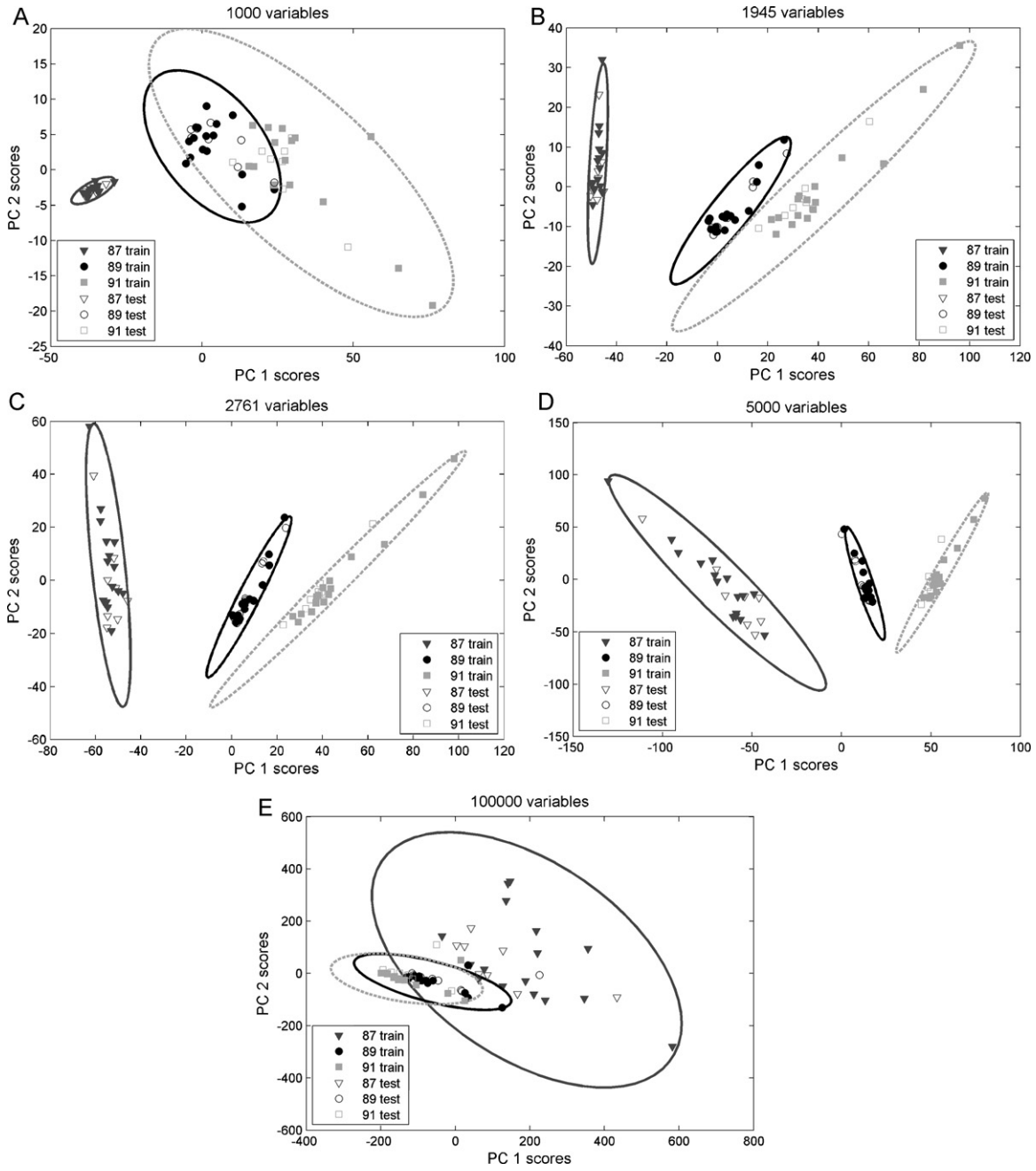


**Fig. 3.** Scores plots for selected PCA models. Dark grey triangles represent 87-octane gasoline, black circles represent 89-octane gasoline and light grey squares represent 91-octane gasoline. Filled markers represent samples used for feature selection and model construction. Hollow markers represent test data to which model was applied. 95% confidence ellipses indicated for each class. A, B, C, D, and E show plots for 1000, 1945, 2761, 5000, and 100 000 included variables, respectively.

the optimum. Then progressively smaller step sizes can be used in the vicinity of the optimum to locate its exact position. Additionally, smaller step sizes must be used as the confidence limit approaches 100% as a relatively small change in confidence limit will result in large change in the size of the confidence ellipses.

Fig. 2 depicts the results of the optimization process for the first of the five sets of data. The cluster resolution between pairs of ellipses is plotted on the *y*-axis vs. the number of features that are included in the model. It is apparent from Fig. 2 that with few variables it is relatively easy to model the differences between 87- and 89-octane gasolines and 87- and 91-octane gasolines. Conversely, it is difficult to distinguish between 89- and 91-octane gasolines, which represent the critical pair of clusters in this case. This figure also highlights the advantage of using a metric that is bounded between 0 and 1. Overall model quality is assessed by taking the product of individual cluster resolutions. In Fig. 2, it is apparent that at a low number of included variables, 89- and 91-octane gasolines are not separated, a fact that is accurately reflected by the product of individual cluster resolutions.

As the number of variables increases, the separation between 89- and 91-octane gasolines shows marked improvement, with 95% confidence ellipses becoming separated when 1945 variables are used and reaching a maximum at 2761 variables. As this pair of clusters was always limiting the quality of the model, the optimal number of features was determined based on the resolution of this pair. Investigating the trend in resolutions past this optimum, a gradual decrease in the resolution for the critical pair is observed until about 20 000 variables (Fig. 2B). Fig. 2C demonstrates the extreme degradation in resolution that is observed when all variables are considered in the model.

**Table 1**
Numbers of variables identified as optimum for a given set using a given metric as well as false positive and false negative rates.

| Set | Cluster resolution | | | | | | Euclidian distance | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Critical pair | | | Product | | | Critical pair | | | Product | | |
| | *n* | FP | FN | *n* | FP | FN | *n* | FP | FN | *n* | FP | FN |
| 1 | 2761 | 0 | 1 | 2761 | 0 | 1 | 29 900 | 8 | 3 | 97 600 | 12 | 2 |
| 2 | 2461 | 0 | 1 | 2461 | 0 | 1 | 9300 | 0 | 4 | 66 600 | 8 | 2 |
| 3 | 2265 | 0 | 1 | 2265 | 0 | 1 | 24 900 | 0 | 1 | 56 800 | 10 | 0 |
| 4 | 2985 | 0 | 1 | 2657 | 0 | 2 | 31 800 | 12 | 2 | 98 400 | 16 | 2 |
| 5 | 3189 | 0 | 2 | 3194 | 0 | 2 | 31 100 | 6 | 10 | 92 600 | 12 | 11 |
| Average | 2732 ± 376 | | | 2668 ± 350 | | | 25 400 ± 9400 | | | 82 400 ± 19 300 | | |
| All train | 2027 | | | 2027 | | | 22 900 | | | 62 000 | | |

FP – false positive. Samples in the test set that do not fall within 95% confidence ellipse of their class. FN – false negative. Samples in the test set that fall inside 95% confidence ellipse of at least one other class. *n* – number of variables at optimum. All Train – numbers of variables at optimum of a set that includes all data.

Fig. 3 depicts the scores plots from the 2-component PCA models constructed using different numbers of variables to highlight regions in Fig. 2. As predicted by the plot in Fig. 2A, a 1000-variable model does not include sufficient features to separate all of the classes and should show overlap of the 89- and 91-octane gasoline samples, while the 87-octane gasoline should be well separated from the other two. The model constructed using 1945 variables should show that all ellipses are just separated at the 95% confidence level, and the model that is constructed using 2761 variables (Fig. 3C) exhibits the best overall resolution between the three classes. The inclusion of additional variables (e.g. the 5000-variable model shown in Fig. 3D) decreases the resolution. In the extreme case when far too many variables are included (100 000, Fig. 3E) the quality of the model is highly degraded, with all ellipses exhibiting significant overlap.

The model constructed using the optimum number of points from the training set, as seen in Fig. 3C, was able to correctly classify the samples in the test set. Similar results have been observed for other sets, as summarized in Table 1. Additionally, when different training sets were selected from the original data set, there was very little difference in the number of variables required to reach the optimum. Moreover, the optimum that is indicated by the least-separated class is identical (or nearly identical) to the optimum determined from the product of the resolutions between all pairings, and in both cases very low error rates are seen at the 95% confidence level (Table 1).

Once important variables have been identified, their positions in the original data can be identified and a binary mask may be generated to visualize the relevant chromatographic information. In the mask, included variables are assigned a value of one and excluded variables are assigned a value of zero. Applying this mask to a chromatogram will only permit the relevant variables to be seen. A mask for the variables included at maximal cluster resolution is presented in Fig. 4. To visualize the order in which variables were added to the model, the mask has been colour-coded according to *F* ratio. Variables with a high *F* ratio are coloured red, variables with a relatively low *F* ratio are coloured green, and variables which are ignored are white. As can be seen, there are two coeluting compounds which are used to discriminate between the three classes of gasoline. Investigation of the raw GC–MS data indicate that the compound responsible for the ions coloured red (ions 91, 78, 65, 52, 39) is toluene. These variables are added first, and as can be seen in Fig. 2A, it is relatively easy to distinguish between 87- and 89-octane and 87- and 91-octane gasoline but difficult to distinguish between 89- and 91-octane gasolines. This is indeed what is observed in the GC–MS data. The 89- and 91-octane gasolines have similarly high concentrations of toluene and other aromatics while the concentrations of these compounds are relatively low in the 87-octane gasoline.

The second compound that was included by the algorithm, indicated by the mostly green points that elute slightly before toluene (ions 43, 57, 71, 85, 99) is a hydrocarbon that coelutes with toluene. Based on the mass spectrum of the compound it is a branched, saturated hydrocarbon having eight carbons, possibly 4-methyl heptane. Inspection of the chromatographic data shows that these features are in fact due to a compound which has a relatively high concentration in the 87- and 89-octane gasoline samples and a relatively low concentration in the 91-octane gasoline. This indicates

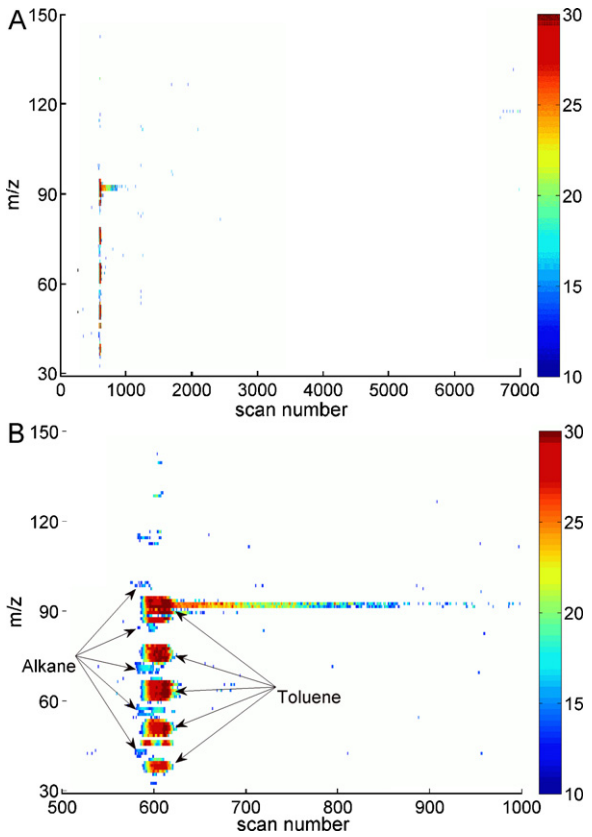


**Fig. 4.** Features in data having the highest 2500 *F* ratios. These features were automatically selected by the data for providing maximal overall cluster resolution. Features have been re-wrapped to indicate the two compounds (toluene and an alkane) in the samples that are used for distinguishing between octane ratings. Features have been colour-coded according to *F* ratio. Red points are high *F* ratio and considered first by the feature selection routine used here, green points are lower *F* ratios and thus considered after the red points. Ignored variables are white. A represents an entire chromatogram worth of data; B is a close up of the region containing the features of interest. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of the article.)

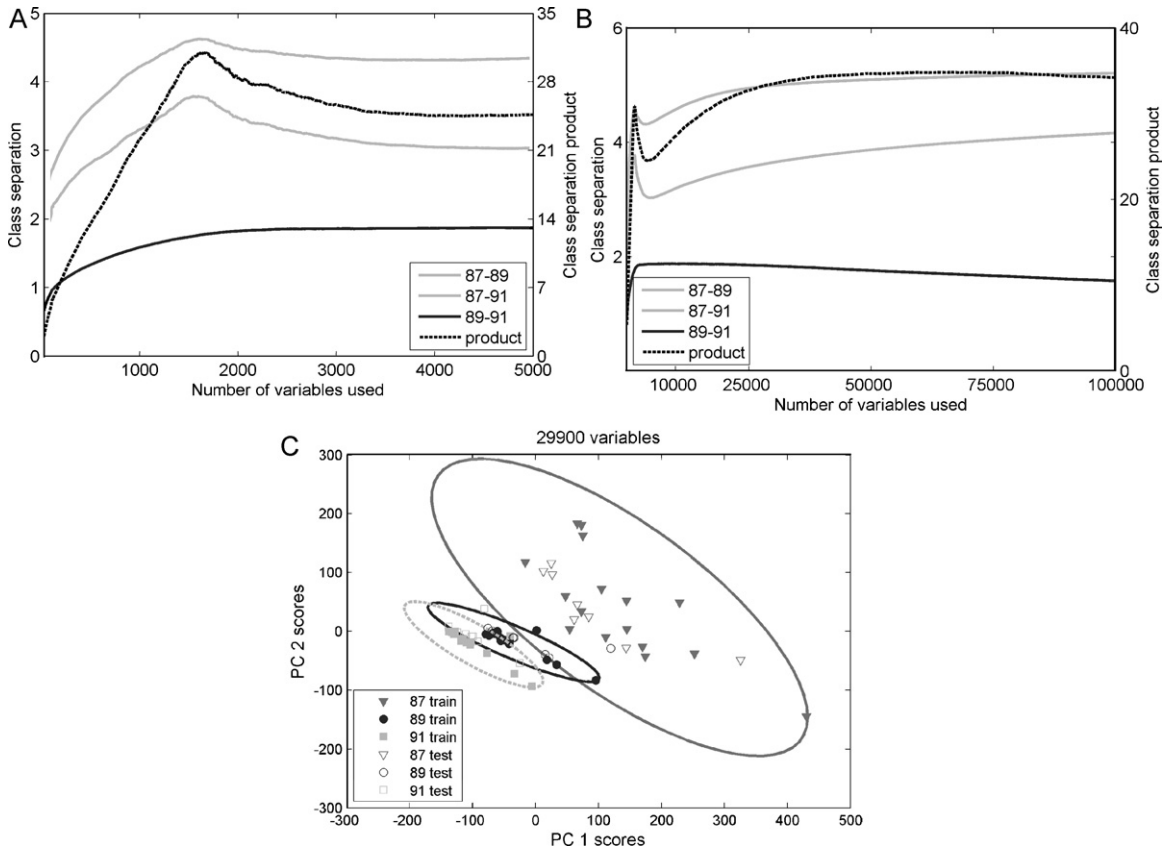


**Fig. 5.** Degree of separation vs. number of variables when calculated using a Euclidean distance approach as well as the product of degree of separation for the three classes. (A) Close up of the region from 0 to 5000 included variables; (B) close up of the region from 0 to 100 000 included variables; (C) scores plot from PCA model using optimal number of variables found in A.

that the model is automatically identifying features in the data that have an actual chemical origin. Furthermore, it shows the power of using the raw chromatographic data over integrated peak tables. If integrated peak areas were used, it is very likely that this compound would not have been observed due to the coelution with the much larger toluene peak.

The other observation is that a further data preprocessing step would ideally be implemented reduce the number of ions considered for the model. Here, ~2500 variables were used to describe essentially two chromatographic peaks. Incorporating the ability to reduce this number of variables to a handful of important ions picked out of key locations in the data is the focus of ongoing efforts.

The cluster resolution metric was compared to a previously described metric based on the Euclidean distance between the centroids of pairs of classes, relative to the square root of the sum of the variance within each group [18,23]. Fig. 5A and B show the degree of class separation calculated based on the Euclidean distance metric for each set of classes, as well as the product of class separations (which has been suggested as a parameter for optimizing overall class separation), using the same test data (Set 1). When the least-separated pair of classes is considered, the optimal separation was observed at 29 900 variables. A visual inspection of the scores plot in Fig. 5C and the one created using the optimum number of variables predicted by the algorithm using cluster resolution (Fig. 3C) shows that the cluster resolution metric provides a model with a significantly more distinct class separation. When product of class separations is considered, the optimum is found to be at 97 600 variables and the model performs similar to the case shown in Fig. 3E (i.e. it fails almost entirely).

Table 1 shows optimal numbers of variables for both metrics for all sets using both the critical pair of classes and the product of all cluster resolutions as the optimization parameter. It additionally indicates the false positive and false negative rates for each training set. It is evident that the cluster resolution metric is far more stable and that the optimum numbers of variables indicated by this metric produces significantly improved models that the Euclidean distance metric.

Additionally, when the two metrics are compared using all 72 chromatograms to train the model, the cluster resolution metric provides its optimum at approximately the same location (2027 variables vs. the average of 2732 variables). Conversely, the Euclidean distance metric reaches its optimum at 25 400 variables, which is clearly too many. When the product of class separations is considered, the ED metric suggests 62 000, over an order of magnitude more than the optimum found using the cluster resolution metric. The reason for this difference is that the new metric considers simultaneously the shapes, sizes, and relative alignments of clusters and is scaled between 0 and 1, so the determining factor will be the resolution of the critical pair of ellipses.

## 5. Conclusions

The cluster resolution metric presented herein provides a means by which the degree of separation between classes of samples can be objectively and automatically quantified in a manner that accounts for the sizes, positions, and relative alignments of the clusters. Additionally, the metric is bounded between 0 and 1, providing a definite advantage for considering overall model quality when more than two clusters of points are considered. This is a

significant step forward in the development of automated feature selection and automated chemometric model development routines. A manner in which this metric can be incorporated into a routine to guide feature selection has also been presented. This metric and general approach to variable selection can in principle be applied to numerous combinations of chemometric models and feature ranking/selection approaches.

Though demonstrated here as a calculation of two-dimensional ellipses on a two-dimensional plane, it is conceivable that the approach can be extended for use in considering the separation between clusters in higher-dimensional spaces. Finally, it should be noted that the analyst must consider the limitations arising from the data or feature ranking technique used when employing cluster resolution to automate feature selection. Splitting data into test sets and training sets for cross-validation is, as always, recommended.

## Acknowledgements

The authors wish 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. The Alberta Ingenuity Fund's support through a New Faculty Award to Dr. Harynuk is also acknowledged. Eigenvector Research, Inc. is acknowledged for discussions and insight into the calculation of confidence ellipses.

## References

[1] P. Doble, P.M.L. Sandercock, E. Du Pasquier, P. Petocz, C. Roux, M. Dawson, Forensic Sci. Int. 132 (2003) 26–39.
[2] P.M.L. Sandercock, E. Du Pasquier, Forensic Sci. Int. 134 (2003) 1–10.
[3] P.M.L. Sandercock, E. Du Pasquier, Forensic Sci. Int. 140 (2004) 43–59.
[4] P.M.L. Sandercock, E. Du Pasquier, Forensic Sci. Int. 140 (2004) 71–77.
[5] I.D. Wilson, R. Plumb, J. Granger, H. Major, R. Williams, E.M. Lenz, J. Chromatogr. B 817 (2004) 67–76.
[6] S.J. Bruce, P. Johnsson, H. Antti, O. Cloarec, J. Trygg, S.L. Marklund, T. Moritz, Anal. Biochem. 372 (2008) 237–249.
[7] U. Lutz, R.W. Lutz, W.K. Lutz, Anal. Chem. 78 (2006) 4564–4571.
[8] T. Kind, V. Tolstikov, O. Fiehn, R.H. Weiss, Anal. Biochem. 363 (2007) 185–195.
[9] J. Vial, H. Noçairi, P. Sassiat, S. Mallipatu, G. Cognon, D. Thiébaut, B. Teillet, D.N. Rutledge, J. Chromatogr. A 1216 (2009) 2866–2872.
[10] 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.
[11] R.E. Mohler, K.M. Dombek, J.C. Hoggard, K.M. Pierce, E.T. Young, R.E. Synovec, Analyst 132 (2007) 756–767.
[12] R. t'Kindt, K. Morreel, D. Deforce, W. Boerjan, J. van Bocxlaer, J. Chromatogr. B 877 (2009) 3572–3580.
[13] B.T. Weldegergis, A.M. Crouch, J. Agric. Food Chem. 56 (2008) 10225–10236.
[14] R.B. Gaines, G.J. Hall, G.S. Frysinger, W.R. Gronlund, K.L. Juaire, Environ. Forens. 7 (2006) 77–87.
[15] C.R. Borges, Anal. Chem. 79 (2007) 4805–4813.
[16] L.J. Marshall, J.W. McIlroy, V.L. McGuffin, R. Waddell Smith, Bioanal. Anal. Chem. 394 (2009) 2049–2059.
[17] D. Ballabio, T. Skov, R. Leardi, R. Bro, J. Chemom. 22 (2008) 457–463.
[18] K.J. Johnson, R.E. Synovec, Chemom. Intell. Lab. Syst. 60 (2002) 225–237.
[19] J.H. Christensen, G. Tomasi, J. Chromatogr. A 1169 (2007) 1–22.
[20] M.D. Krebs, R.D. Tingley, J.E. Zeskind, M.E. Holomboe, J. Kang, C.E. Davis, Chemom. Intell. Lab. Syst. 81 (2006) 74–81.
[21] T. Rajalahti, R. Arneberg, A.C. Kroksveen, M. Berle, K.M. Myhr, O.M. Kvalheim, Anal. Chem. 81 (2009) 2581–2590.
[22] N.E. Watson, M.M. VanWingerden, K.M. Pierce, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1129 (2006) 111–118.
[23] K.M. Pierce, J.K. Hope, K.J. Johnson, B.W. Wright, R.E. Synovec, J. Chromatogr. A 1096 (2005) 101–110.
[24] R.F. Teófilo, J.P.A. Martins, M.M.C. Ferreira, J. Chemom. 23 (2009) 32–48.
[25] J.H. Christensen, A.B. Hansen, U. Karlson, J. Mortensen, O. Andersen, J. Chromatogr. A 1090 (2005) 133–145.
[26] K.M. Pierce, J.C. Hoggard, J.L. Hope, P.M. Rainey, A.N. Hoofnagle, R.M. Jack, B.W. Wright, R.E. Synovec, Anal. Chem. 78 (2006) 5068–5075.
[27] R.G. Brereton, Applied Chemometrics for Scientists, John Wiley & Sons Inc., Toronto, 2007.
[28] M. Zhang, W. Wang, Y. Du, Chemom. Intell. Lab. Syst. 102 (2010) 84–90.
[29] H. Li, J. Sun, J. Forecast. 29 (2010) 486–501.
[30] S. Wold, Chemom. Intell. Lab. Syst. 2 (1987) 37–52.
[31] K.M. Pierce, J.C. Hoggard, R.E. Mohler, R.E. Synovec, J. Chromatogr. A 1184 (2008) 341–352.
[32] R. De Maesschalck, D. Jouan-Rimbaud, D.L. Massart, Chemom. Intell. Lab. Syst. 50 (2000) 1–18.
[33] M.S. Srivastava, C.G. Khartri, An Introduction to Multivariate Statistics, Elsevier North Holland, Inc., New York, 1979.
[34] G. Almkvist, B. Berndt, Am. Math. Monthly 95 (1988) 585–608.
[35] K.J. Johnson, B.W. Wright, K.H. Jarman, R.E. Synovec, J. Chromatogr. A 996 (2003) 141–155.