frontiers | Frontiers in Analytical Science

REVIEW
published: 19 May 2022
doi: 10.3389/frans.2022.867938

Review of Variable Selection Methods for Discriminant-Type Problems in Chemometrics

*Michael D. Sorochan Armstrong, A. Paulina de la Mata and James J. Harynuk**

Department of Chemistry, Harynuk Research Group, the Metabolomics Innovation Centre, University of Alberta, Edmonton, AB, Canada

Discriminant-type analyses arise from the need to classify samples based on their measured characteristics (variables), usually with respect to some observable property. In the case of samples that are difficult to obtain, or using advanced instrumentation, it is very common to encounter situations with many more measured characteristics than samples. The method of Partial Least Squares Regression (PLS-R), and its variant for discriminant-type analyses (PLS-DA) are among the most ubiquitous of these tools. PLS utilises a rank-deficient method to solve the inverse least-squares problem in a way that maximises the co-variance between the known properties of the samples (commonly referred to as the *Y*-Block), and their measured characteristics (the *X*-block). A relatively small subset of highly co-variate variables are weighted more strongly than those that are poorly co-variate, in such a way that an ill-posed matrix inverse problem is circumvented. Feature selection is another common way of reducing the dimensionality of the data to a relatively small, robust subset of variables for use in subsequent modelling. The utility of these features can be inferred and tested any number of ways, this are the subject of this review.

Keywords: chemometrics, classification, metabolomics, mass spectrometry, spectroscopy, chromatography, machine learning, variable/feature selection

OPEN ACCESS

Edited by:
Raffaele Vitale,
Université de Lille, France

Reviewed by:
Federico Marini,
Sapienza University of Rome, Italy
Rosalba Calvini,
University of Modena and Reggio Emilia, Italy

****Correspondence:***
James J. Harynuk
james.harynuk@ualberta.ca

Specialty section:
This article was submitted to Chemometrics, a section of the journal Frontiers in Analytical Science

Received: *01 February 2022*
Accepted: *21 April 2022*
Published: *19 May 2022*

Citation:
Sorochan Armstrong MD, de la Mata AP and Harynuk JJ (2022) Review of Variable Selection Methods for Discriminant-Type Problems in Chemometrics. Front. Anal. Sci. 2:867938. doi: 10.3389/frans.2022.867938

1 INTRODUCTION

The output of modern chemical instrumentation can provide a high degree of dimensionality, or number of features, to describe each sample that can be used to gain insight into complex chemical mixtures. Analysts often use this information to construct models for discriminant-type problems, but the burden of dimensionality limits the application of typical algorithms such as Support Vector Machines (SVM) (Crammer and Singer, 2001), or Canonical Variates Analysis (CVA). CVA is commonly referred to in the literature as: Linear Discriminant Analysis (LDA) (Nørgaard et al., 2006), and the two terms are often used interchangeably. However LDA refers specifically to the calculation of one or many classification thresholds while CVA refers to the dimensionality reduction technique that maximises class separation along the latent variables. Reducing the dimensionality is critical for representative and interpretable models, and often relies on methods such as Partial Least Squares (PLS) (Wold et al., 2001) to weight variables in the *X*-block that are highly co-variate with those values in the *Y*-block. However, the resultant regression coefficient output is not always particularly informative, since the analyst needs to make decisions for which regression coefficient loadings are more or less significant than others for inclusion into their resultant interpretation of the data. PLS-DA alone offers very little recourse in those instances where the regression coefficient scores of the models are not particularly useful for classifying samples external to the training set.

Applications of discriminant-type problems have enjoyed considerable attention over the past few decades, thanks in part to the explosion of interest in metabolomics (Dettmer et al., 2007). By examining the differences in metabolite expression, quantitative differences between biological states can be inferred through the use of discriminant analyses. This can simplify the problem of biological interpretation, but the question of what regression coefficient scores ought to be deemed significant is still relevant. Simplifying the output of the data analysis routine to include only those features that are deemed to be significant for subsequent interpretation is one motivation for employing a feature selection routine.

There is an idealised linear relationship between the analytical response of one or several chemical factors and their absolute quantities. While this relationship is not always observed at the limits of the dynamic range of the instrument, through careful method optimisation it is usually safe to make the assumption that the underlying chemical phenomena can be studied using linear methods. This is as true for regression-type problems as it is discriminant-type problems, and simplifies the practical application of these technologies. Hence in chemometrics the focus for variable selection and modelling typically favours linear methods (Hopke, 2003). Though better performance has in some cases, been reported for non-linear models (de Andrade et al., 2020), linear models are more easily interpretable and thus favoured for applications where underlying correlation and causal relationships are being sought.

Different assumptions can be made about the data, depending on whether a discriminant or regression analysis is being performed, and variable selection techniques that may be appropriate for regression-type problems may not be applicable for discriminant-type problems. Known characteristics of the *X*-block are also relevant for certain variable selection routines, based on the properties of the instrumentation used for data collection. In this article, we distinguish between two basic types of data, based on what are commonly encountered by analysts: discrete, pre-processed, identifiable, tabulated data (e.g.,: peak table data output from a chromatographic system), and continuous data (e.g.,: raw chromatographic or spectroscopic data). More tools can be applied for continuous data, since "windows" of adjacent variables can be considered all at once and variables within certain regions can be assumed to correlate with one another and some underlying chemical information. However, variables will often correlate to multiple chemical species simultaneously, making interpretation less straightforward.

1.1 Discriminant Analyses as Regression

Feature selection may be used to improve the mathematical characteristics of the problem–to avoid the computation of an ill-poised problem. For either regression or discriminant-type problems, when the number of variables exceeds the number of observations, or samples, there are an infinite number of possible solutions to the equation:

$$Y = X\beta \tag{1}$$

The solution to **Eq. 1** identifies a useful variable subset in X that is able to accurately predict those values in Y via the regression coefficient, β. This in effect is minimising the following cost function in the least-squares sense, yielding an solution to **Eq. 1** via β

$$SSR = \|Y - X\beta\|_2^2 \tag{2}$$

$$\hat{\beta} = argmin_{\beta}(SSR) = \left(X^T X\right)^{-1} X^T Y \tag{3}$$

Where $\|.\|_2^2$ denotes square of the L_2 norm of error term, often referred to as the sum of squared residuals (SSR) and $\hat{\beta}$ is the predicted value for β, determined empirically in a way that minimises **Eq. 2**.

In this case, the only meaningful difference between a discriminant-type analysis and a regression-type analysis is the content of the *Y*-block. Namely, if the *Y*-block contains categorical information for two or more classes, then the information encoded is discrete, and the analysis is a discriminant-type analysis. For continuous, quantitative information encoded in the *Y* block, the analysis is generally referred to as a regression. Considerations for scaling are also different for regression-type vs discriminant-type problems: while scaling is critical for a *Y*-block of continuous data, it is optional for categorical data since the manner in which the observations are weighted is consistent across all observations. However, treatment of a classification problem as a regression problem is only commonly seen in those instances where PLS is used, although principal component scores have also been investigated (Yendle and MacFie, 1989). Even in those cases where PLS-DA is employed, a number of critical parameters such as residuals must be ignored, since a line-of-best-fit through binary classification data is assumed to have high and poorly informative residuals due to the very nature of regression problems. Within the context of a regression, a feature selection routine indicates which variable contributions are negligible, and removes them from the model. In essence, this is analogous to setting certain variable contributions within a regression vector to zero.

1.2 Canonical Variates Analysis (CVA)

Rather than calculating a linear model that best minimises the sum of squared residuals through categorical data, it is also possible to deploy CVA to maximise Fisher's discriminant ratio (FDR) (similar to **Equation 8**) of two or more classes based on each sample's projection scores on a series of latent variables. This method is more robust against outliers and unequal numbers of samples versus regression methods, but assumes homoscedasticity for deriving the decision threshold (Theodoridis, 2020). CVA is calculated via the co-variance matrices that reflect within-class variance, versus between-class variance Nørgaard et al. (2006):

$$S_{within} = \frac{1}{(n-g)} \sum_{i=1}^{g} \sum_{j=1}^{n_j} \left(\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i}\right)\left(\mathbf{x_{ij}} - \bar{\mathbf{x}}_\mathbf{i}\right)^T \tag{4}$$

$$S_{between} = \frac{1}{(g-1)} \sum_{i=1}^{g} n_i \left(\bar{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}}\right)\left(\bar{\mathbf{x}}_\mathbf{i} - \bar{\mathbf{x}}\right)^T \tag{5}$$

$\mathbf{x_{ij}}$ refers to the i^{th} class and j^{th} observation, and n_i and n_j refer to the total number of classes and observations, respectively. For S_{within}, each entry in $\mathbf{x_{ij}}$ is scaled relative to the i^{th} class, denoted as $\overline{\mathbf{x_i}}$. For $S_{between}$, the variance of each class centroid, $\overline{\mathbf{x_i}}$ is determined relative to the overall mean, $\bar{\mathbf{x}}$.

A solution for CVA can be found once the problem is written as an eigenvalue problem:

$$S_{within}^{-1} S_{between} \mathbf{w} = \lambda \mathbf{w} \tag{6}$$

Where $\mathbf{w}$ is the eigenvector, or latent variable that maximises $F(\mathbf{w})$:

$$F(\mathbf{w}) = \frac{\mathbf{w}^T S_{between} \mathbf{w}}{\mathbf{w}^T S_{within} \mathbf{w}} \tag{7}$$

As a classifier, CVA is used relatively infrequently relative to PLS-DA, since it cannot handle singular co-variance matrices as either S_{within} or $S_{between}$. However, since PLS-DA makes use of a regression to effect its discrimination, it does suffer from the same drawbacks as any other classifier that is informed by a linear line-of-best-fit. CVA has been modified to account for ill-posed problems where PLS-DA is typically applied, through its application as a technique for solving the matrix inverse of high-dimensional data. This was first described by (Nørgaard et al., 2006) as Extended Canonical Variates Analysis (ECVA). For either traditional CVA, or ECVA the decision threshold is calculated by fitting a multivariate normal distribution to each class, and assigning predictions based on each sample's highest modelled likelihood of belonging to each class.

1.3 Types of Feature Selection

Feature selection routines can be categorised as belonging to either filter, wrapper, or embedded methods (Kohavi and John, 1997). There also exist hybrid methods that combine at least two of these approaches. True to their name, filter methods use some variable ranking scheme, and include only those variables with a value greater than a particular threshold. While simple and computationally efficient, filter methods require extensive user intervention to determine an appropriate value for the threshold, and may or may not account for correlations between variables depending on the variable ranking metric used. Wrapper methods evaluate several candidate variable subsets and select the variable subsets with the best performance. While wrapper methods can be robust, they are computationally expensive due to a inherent degree of redundancy built into the algorithms. Embedded methods include variable selection as a part of the calculation of the model itself.

This review will describe methods for variable selection, rather than methods for variable weighting such as Projection Pursuit Analysis (PPA) (Hou and Wentzell, 2011), PLS, or manifold learning for non-linear modelling (Van der Maaten and Hinton, 2008). While variable significance can be judged and subsequent variable selection performed this requires human intervention, and so such techniques are widely considered to be dimensionality reduction techniques rather than variable selection techniques.

2 FILTER METHODS

Filter methods proceed following variable ranking, and may require user intervention in order to determine an appropriate cut-off for variable significance. Variables scoring above the threshold are included in the model, and the rest are discarded. The advantage of filter methods is that they are relatively easy to apply, and certain variable ranking metrics may include the importance of co-variate or correlated variables as they affect a latent variable discriminant analysis such as PLS-DA or CVA (Kvalheim, 2020). However since many latent variable discriminant analyses have a tendency to over-fit the data, the use of variables acquired using metrics derived from the latent variable model are only as useful as the latent variable model itself.

2.1 Variable Ranking Metrics

2.1.1 Fisher(F)–Ratios

The simplest variable ranking filter is arguably the Fisher or *F*-ratio, which describes the pooled variance of all samples over the variance attributable to class means relative to the overall mean (Maddala and Lahiri, 1992). These two considerations have already been described in **Eq 4**, **Eq. 5**, and the *F*-ratio is described quite simply as:

$$F = \frac{S_{between}}{S_{within}} \tag{8}$$

A key difference between the *F*-ratio for variable significance, versus for CVA, is that the F ratio is calculated for individual variables, and a multivariate optimisation for class separation is not performed.

The *F*-ratio described by the *F*-distribution, which includes degree-of-freedom values for both the numerator and denominator as $(g - 1)$ and $(n - g)$ respectively as input parameters, where n is the total number of samples and g refers to the number of classes. Critical values of significance can be used to inform the appropriate threshold based on the parametric *F*-Distribution, but there is no guarantee that the observed distribution of *F*-ratios will follow a theoretical one - especially in cases where the data is not normally distributed, since **Equation 8** can also be written as:

$$F(\nu_1, \nu_2) = \frac{\chi_1^2 / \nu_1}{\chi_2^2 / \nu_2} \tag{9}$$

Where the numerator and denominator are χ^2 values with degrees of freedom ν_1 and ν_2 respectively. This relationship highlights the Gaussian assumptions of the *F*-distribution (Box, 1953), and by extension CVA.

The *F*-ratio has been used extensively in discriminant-type analyses of chemical data, owing to the simplicity of its application, relationship to one-way ANOVA via **Eq. 9** and the fact that direct variable-variable comparisons evade issues surrounding the dimensionality of the data. Experiments where the number of identifiable underlying chemical features greatly outnumber the number of samples (e.g.,: many omics problems

and non-target environmental analyses) see frequent use of the *F*-ratio (Johnson and Synovec, 2002; Marney et al., 2013; de la Mata et al., 2017; Pesesse et al., 2019), in part because the costs associated with sample analysis make any attempts to lessen the impact of dimensionality much less practical. Since the objective of any analysis of *F*-ratios is a descriptive rather than an inferential one, it could be argued that *F*-ratio measurements are not a learning technique, but rather a descriptive statistical analysis. However, the act of applying a threshold for variable significance introduces bias to the model which nonetheless warrants further confirmation with external samples.

F-ratios can be applied on either continuous or peak-table data. Synovec's research group (Pierce et al., 2006; Marney et al., 2013) has frequently compared *F*-ratios of each ion channel (mass-to-charge ratio, m/z) from samples of known classes to qualitatively identify regions of two-dimensional chromatograms where there is "discriminating" information using either a pixel-based or tile-based approach. In these approaches, the variance across individual mass channels are compared via Fisher ratio, accounting for known class membership, and the Fisher ratios are summed across all mass channels to visualise the significance of each pixel on a chromatogram. It is worth noting that the pixel-based approach can falsely indicate significance if there is significant chromatographic drift between samples. This issue has been overcome somewhat by the tile-based approach, wherein peaks are expected to drift within the tile spaces themselves. In the tile-based approach, an additional summation step is used to indicate the quantity of a particular ion within the tile. For complex data and/or large studies with hundreds or thousands of samples, there may still be issues with peaks drifting in and out of tiled regions across the data set if the tiles are not sized appropriately or if signals drift too much over the course of the study (this would affect the subsequent determination of the *F*-ratios).

2.1.2 Selectivity Ratios

The Selectivity Ratio (SR) (Rajalahti et al., 2009) is another metric for variable ranking. It encodes multivariate and co-linearity information within the rank of each variable as informed by the ratio of its variance explained within the predictive model versus its variance within the residual matrix:

$$X_{LV} = TP^T + E \tag{10}$$

Where X_{LV} is an $m \times n$ reconstruction of observations and variables via a latent variable analysis such via PLS, as informed by a regression model $Y = Xb$, and E is an error or residual matrix of similar dimensions. SR is defined for a particular variable as the ratio between the variance explained in the latent variable space versus its contributions to the noise. Here, for $j \in [1, n]$, SR_j is the selectivity ratio for the j^{th} variable out of n:

$$SR_j = \frac{\|TP_j^T\|^2}{\|E_j\|^2} \tag{11}$$

This method has been applied in a number of studies (Rajalahti et al., 2009; Amante et al., 2019) for discriminant-type problems, due in part to the ease with which it can be integrated within the framework of PLS-DA.

2.1.3 VIP Scores

Variable Importance in Projection (VIP) scores are a measure of a particular variable's influence on the latent variable model, which is correlated to the variance explained in the *Y* block. They are calculated as the weighted sum of squares as a product of the amount of variance explained by the model (Farrés et al., 2015). VIP scores were originally described by (Wold et al., 1993). However, the article (Chong and June 2005) arguably first popularised the method, which is described using **Equation 12**:

$$VIP_j = \sqrt{\frac{n\sum_{i=1}^{k} b_i^2 \mathbf{t_i}^T \mathbf{t_i} \left(\mathbf{w_{ij}} / \|\mathbf{w_i}\|\right)^2}{\sum_{i=1}^{k} b_i^2 \mathbf{t_i}^T \mathbf{t_i}}} \tag{12}$$

In **Equation 12**, for a matrix of $m \times n$ samples by variables, VIP_j refers to the VIP score for the j^{th} variable, and n refers to the total number of variables. For a k-component PLS model, b_i indicates the i^{th} entry of the regression vector for the i^{th} component derived from the vector or matrix of observed values relative to the score matrix, T. $\mathbf{t_i}$ indicate the i^{th} vector of the score matrix T. $\mathbf{w_{ij}}$ indicate the weights of the PLS model in the X block for each component, i, and variable j normalised to the euclidean norm of the weight vector for the particular component.

Many analysts use a threshold of 1 (Stoessel et al., 2018) to indicate what VIP scores are significant, but examination of the resultant projections before and after variable inclusion or exclusion are advised to prevent over-fitting of the data (Andersen and Bro, 2010). VIP scores are included as a method of variable ranking in the popular online platform MetaboAnalyst (Pang et al., 2021), and are used in many publications, with an unsurprisingly high representation in fields related to metabolomics (Seijo et al., 2013; Stoessel et al., 2018; Ghisoni et al., 2020; Sinclair et al., 2021).

Although other variable ranking metrics are used (Tran et al., 2014; Mehmood et al., 2020), the *F*-ratio, selectivity ratio, and VIP score are among those most frequently encountered in chemometrics for discriminant problems. Selectivity ratios and VIP scores can tend towards over-fitting the data, since they rely on parameters returned by PLS-DA which suffers from problems both related to dimensionality and the application of a regression model for a discriminant-type problem. Despite this, selectivity and VIP scores account for co-linearity, unlike the *F*-ratio. Despite this drawback, the calculation of *F*-ratios is a more statistically informative criterion for discriminant-type problems than VIP scores and the selectivity ratio, and remains popular as a simple method for selecting variables that feature a high degree of univariate discrimination.

Also of note: Talukdar et al. (2018) utilised non-linear kernel partial least squares (kPLS) and eliminated poorly weighted coefficients in the model to improve prediction accuracy, indicating that filter methods can be easily applied in conjunction with more sophisticated modelling and variable significance techniques.

3 WRAPPER METHODS

Wrapper methods evaluate a number of different variable subsets through multiple iterations of the algorithm, and return the best-performing variable subset Kohavi and John (1997). Choice of performance metrics used to evaluated the variable subsets, as well as the methods for determining the variable subsets have a profound effect on the resultant output of the algorithm. Wrappers can evaluate a single operation iteratively through a single dataset (Rinnan et al., 2014) followed by validation on an external or cross-validated set, or can be evaluated at each iteration using different combinations of the data (Sinkov et al., 2011).

3.1 Variable Subset Selection

3.1.1 Forward Selection, Backwards Elimination

Regardless of the method used to determine a variable subset, some initial assumptions about what variables are likely to be useful in the final model must be made, since generating and evaluating random subsets is computationally prohibitive for high-dimensional data. This may qualify certain wrapper methods as hybrid approaches, due to their reliance on a filter method in a preliminary variable ranking step. The simplest methods for variable subset selection use either forward selection, where variables are successively added to a model; or backwards elimination, where variables are successively removed from a model. In either case, the performance of the variable subset is determined at each step where a variable is either added or removed. Studies that use a hybrid forward selection/backwards elimination routine (Sinkov et al., 2011) have been proposed for systems when there are many thousands, or even millions of data points describing each sample. When using one of either forward selection or backwards elimination, the only challenge is selecting a high-performing variable subset, and frequently all variables are tested. But in order to utilise both forward selection and backwards elimination, an initial population of variables for the backwards elimination must be chosen, followed possibly by a point where one should stop the subsequent forward selection. This has been done by estimating the distributions of each variable's Fisher ratios (Adutwum et al., 2017), and an estimate for the optimal "start" and "stop" was performed within this framework using numerical experiments.

3.1.2 Genetic Algorithms

Genetic algorithms (GAs) are another method for selecting a variable subset: high performing variable subsets "evolve" and are selected to share information with other informative subsets to create new subsets. The theory being that at each iteration the surviving subsets become more informative as variables contained within poorly performing subsets are excluded from further consideration. Genetic algorithms are poorly characterised mathematically, since they imitate biological processes, rather than try to exploit specific mathematical characteristics of the data. They are best understood through their dynamic programming routine, which can be summarised for classification problems (Cocchi et al., 2018):

Algorithm 3.1.2: Variable Subset Selection by Genetic Algorithm.

1. A user-defined (A) number of individual "chromosomes", each containing binary information indicating the presence or absence of the variables, are randomly generated and the performance criteria for each evaluated.
2. A user-defined (B) number of highly performing individual chromosomes are selected to move forward to the next "generation".
3. The highly performing individual chromosomes randomly exchange information to create new individuals, typically of the same population size (A). In each case there is a small probability (C) for a random mutation to occur. Following this, the fitness of each individual chromosome is reassessed.
4. Reiterate steps 2, 3 for a set number of iterations (D_a), or until a performance criterion D_b is reached (Lavine et al., 2011).

As described in Algorithm 3.1.2, there are a number of required user input parameters that can have a profound impact on the feature selection routine. Genetic algorithms can also be quite slow due to their reliance on dynamic programming to select a high-performance variable subset. As such, the use of GAs can be difficult for high-dimensional data. Ballabio et al. (2008) demonstrated dimensionality reduction techniques prior to a feature selection step using GAs, to circumvent this problem.

Lavine has proposed a number of feature selection methods based on GAs acting as wrapper functions, for source identification of jet fuel using Solid Phase Microextraction (SPME) and Gas Chromatography (Lavine et al., 2000) and fuel spill identification (Lavine et al., 2001) among others.

A drawback of wrapper methods, is that while the performance evaluation functions themselves are generalisable, the same cannot be said for their implementation. Oftentimes, several nested internal validation routines are used to train and evaluate the candidate features in a way that is tailor-made to the data, especially if there are hierarchical classifications involved. If these routines are not perfectly transparent they may not be reproducible, and are poorly generalisable to data with different characteristics.

3.1.3 Methods Based on PLS

Many wrapper methods based on PLS exist in the literature - arguably the most widely-known and most influential of these methods is Uninformative Variable Elimination (UVE) (Centner et al., 1996). Through several iterations, a model is trained on a number of samples, and variables that are under-performing (typically based on an analysis of the regression vector) are eliminated. The resultant variable subset is then assessed based on its ability to correctly indicate the samples external to the model, and variables that are consistently selected are included in the final model. While most broadly applicable for regression-type analyses, this technique has also been used to distinguish different cultivars of corn using Terahertz (far-infrared) spectroscopy (Yang et al., 2021).

For spectroscopic data with a high degree of co-linearity, different wavelength bands can be selected as candidate variable subsets. This is the principle of Interval-PLS (Nørgaard et al., 2000). It has also been applied in discriminatory analyses: for example, Peris-Díaz et al. (2018) compared the performance of different spectral regions based on their ability to distinguish different samples of amber based on geological age and region of origin with Raman spectroscopy. The wrapper function in this case is the use of several PLS models calibrated on different spectral regions, which can be either sequentially backwards eliminated or forward selected (Mehmood et al., 2012).

Although a number of methods based on PLS exist, it is also worth noting that not all of these methods have been applied for discriminant-type analyses. Recursive weighted partial least squares (rPLS) (Rinnan et al., 2014), has not been widely demonstrated on discriminant problems, although it has been used in one recent publication to select relevant features for different cultivars of saffron by Aliakbarzadeh et al. (2016). In this publication, the features selected by rPLS appeared to correspond well to features selected using other methods. rPLS iterates through incrementally scaled versions of the X block, weighted using the regression coefficient (for PLSR) calculated at each iteration (r), where D_r is the diagonal of the regression vector (b) that is normalised to its maximum value in **Equation 13**. The algorithm reaches convergence once each variable that is included in the final variable subset approaches 1, and the ones not deemed to be significant approach 0.

$$X_r = X_{r-1} D_r \tag{13}$$

$$X_R = X \prod_{r=1}^{R} D_r \tag{14}$$

3.2 Methods for Evaluating the Performance of the Discriminant Function

3.2.1 Traditional Measures

Performance metrics can strongly inform the resultant output of a wrapper-based method that iterates through several combinations of internal training and test sets. The results of a discriminant type analysis can typically be summarised in the form of a confusion matrix, which is a table that summarises the distribution of the predicted samples classes relative to their known class. Although accuracy (the number of true positives and true negatives over the total number of samples) is a commonly used metric, it is not particularly informative on grossly unbalanced datasets, and is inappropriate in fields such as diagnostics where sensitivity (the number of correctly indicated positive samples over the number of known members of the "true" class) is a much more important consideration than specificity (the number of correctly indicated negative samples over the known number of members of the "negative" class) for binary classification problems. Nonetheless, considerations for the predictive power of a classification model must be appropriately summarised in order to simplify the problem such that the variable subset can be optimised relative to a single cost function.

3.2.2 F_β Scores

The F_1-score (Not to be confused with the F-ratio in this manuscript) is a summary of the model performance, incorporating measures of sensitivity and precision (true positives over the sum of true positives and false positives), and is given by the equation:

$$F_1 = 2\left(\frac{PPV \times TPR}{PPV + TPR}\right) = \frac{2TP}{2TP + FP + FN} \tag{15}$$

Where the F_1 score can be summarised simply as the harmonic mean of sensitivity and precision. In those instances where the analyst would like to bias the performance measure somewhat to better reflect the needs of the analysis, it is also possible to incorporate an additional β coefficient that weights the relative importance of sensitivity vs. precision:

$$F_\beta = \left(1 + \beta^2\right) \frac{PPV \times TPR}{\left(\beta^2 \times PPV\right) + TPR} \tag{16}$$

Where PPV describes precision (Positive Predictive Value) and TPR describes sensitivity (True Positive Rate).

3.2.3 Area Under the Curve

The Area Under the Curve (AUC) value for the Receiver Operator Characteristics of a classification is also used as a measure of model performance. The Receiver Operator Characteristics of a binary classifier plot the True Positive Rate against the False Positive Rate as a function of the position of the decision boundary relative to the samples being classified. Intuitively, a decision boundary that classifies all samples as belonging to class 1 would have both a high false positive and true positive value. If the samples being considered are perfectly resolved along the discriminant axis, then the true positive rate has no effect on the false negative rate. However, for misclassified samples, a reduction in the true positive rate has some demonstrable effect on the false negative rate. The AUC is the area under the Receiver operator curve, which is a value between 0 and 1, with 1 being a perfectly performing classifier. This has not frequently been used to guide feature selection routines in chemometrics, although it has been demonstrated in adjacent fields (Wang and Tang, 2009).

3.2.4 Cluster Resolution

Cluster Resolution (CR) is a statistical measure of model performance that measures the maximum confidence interval over which two confidence ellipses are non-intersecting in two or more dimensions. This has been demonstrated in a linear subspace such as principal component space for uncorrelated, orthogonal scores. It was first described by Sinkov and Harynuk (2011), using a dynamic programming approach. A numerical determination was proposed by Armstrong et al. (2021), that minimised the χ^2 value of significance between a binary set of clusters. It is a particularly useful method for those instances

where there are a number of missing values in the X block, which is a persistent problem in non-target analytical studies involving complex samples. In these data sets, single chemical features are not always reliably registered in the same column across the X-block across all samples, and when considering trace compounds, they may not always be properly detected if their abundance is near the abundance threshold for inclusion in the data table. An advantage of CR is that single chemical components registered in multiple columns of the dataset will be highly correlated, and thus identifiable as being useful features when projected into principal component space. CR is also a less granular metric than those derived from classification performance metrics, and can offer information about the discriminatory power of different variable subsets without samples crossing a decision threshold. Therefore, the observation of relatively minor changes are easily scrutinised when evaluating a feature selection routine. However, a drawback of this technique is that, when compared to other methods, it requires a larger number of samples (20–30 per class as a minimum) for proper estimation of variances along the principal component axes.

4 EMBEDDED METHODS

4.1 Regularisation Methods

Regularisation methods add an additional term to the minimisation of the least-squares problem, that constrains some components of the regression coefficient (β) to be equal to zero. These methods constrain the length of β, via a regularisation constant γ, such that certain entries of the regression vector are weighted more significantly, or with λ such that certain entries are excluded from the final model as zeros.

$$\hat{\beta} = argmin_{\beta}\|Y - X\beta\|_2^2 + \gamma\|\beta\|_2^2 \tag{17}$$

The length of the regression vector can be constrained via the Euclidean norm (the L_2 norm) as in **Equation 17**, where it is described as being a ridge regression or Tikhonov regularisation. Or, via the Lagrangian (L_1) norm:

$$\hat{\beta} = argmin_{\beta}\|Y - X\beta\|_2^2 + \lambda|\beta|_1 \tag{18}$$

The sparse regression coefficient, $\hat{\beta}_R$ for **Equation 17** can be solved analytically as:

$$\hat{\beta}_R = \left(X^TX + \gamma I\right)^{-1}X^TY \tag{19}$$

Equation 18 is referred to as a Least Absolute Shrinkage and Selection Operator (LASSO) Regression by Tibshirani (1996), although a similar approach was first described earlier by Santosa and Symes (1986). The L_2 norm of any vector ($\mathbf{x}$) is can be described as $\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\sum_{i=1}^{n} x_i^2}$, and the L_1 norm as the sum of the absolute values of $\mathbf{x}$: $|\mathbf{x}|_1 = \sum_{i=1}^{n}|x_i|$. In either case, the regularisation coefficient, λ, must be selected by the user. This is typically done by picking a value that maximises the prediction accuracy of the model, achieved by analysing the sum of residual squares for a set of previously unconsidered samples in the case of regression-type problems, or by analysing the prediction accuracy for discriminant-type problems.

However for LASSO, the coefficient $\hat{\beta}_L$ must be determined through convex optimisation. Despite reliance on numerical methods for optimisation, it has been proven that a unique solution exists for LASSO regressions of ill-posed problems (Tibshirani, 2013).

An extension of LASSO for use in discriminant analyses via CVA was proposed by Trendafilov and Jolliffe (2007); however, it was applied on datasets containing relatively few numbers of variables, which is not a common occurrence in chemometrics. Witten and Tibshirani (2011) implemented a different formulation of the problem, and proved its utility on far higher dimensionality data, which has been released as the R package *penalizedLDA* (Witten and Witten, 2015). This package has been used in chemometrics-type work, in distinguishing wild-grown and cultivated *Ganoderma lucidum* using Fourier transform infrared spectroscopy (Zhu and Tan, 2016), which also included a comparison of various other methods for sparse discriminant analyses. Also of note, was LASSO coupled to a logistic regression for classification of different fabric dyes using UV-Vis Spectroscopy (Rich et al., 2020).

For values of $m \ll n$ in X, it may be helpful to select a number of variables that are highly correlated with each other–especially for spectral data where the bandwidth of any particular transition implies a degree of co-linearity within the dataset. Using a LASSO regression, only the most highly correlated variables are usually included in the final model, but due to the regularisation parameter some correlated variables may be lost. As a consequence, the model may be less predictive, and it may be harder to interpret the output of the model. By adding both regularisation parameters from **Eqs 17**, **18**, it is possible to include more correlated features in the final model:

$$\hat{\beta} = argmin_{\beta}\|Y - X\beta\|_2^2 + \gamma\|\beta\|_2^2 + \lambda|\beta|_1 \tag{20}$$

Based on **Equation 20**, Clemmensen et al. (2011) developed a sparse extension of CVA for classification of highly co-linear data, which is of particular interest in chemometrics. This algorithm has been released as the R package: *sparseLDA* (Clemmensen and Kuhn, 2016) and has been used to analyse data from a Direct Analysis in Real Time - High Resolution Mass Spectrometry (DART-HRMS) experiment, to the end of classifying personal lubricants for forensic analysis (Coon et al., 2019), and for distinguishing different printer inks by micro-Raman spectroscopy (Buzzini et al., 2021).

Sparsity has also been used for variable selection via PLS-DA. This approach was first reported by using a sparse PLS regression step as described by Chun and Keleş (2010), followed by classification using a standard classifier such as CVA or a logistic regression. Lê Cao et al. (2011) developed a single-step solution for sparse PLS-DA using an approximation of the LASSO regularisation function via:

$$argmin_{u_i,v_i}\|X^TY - u_iv_i^T\|_F^2 + sign(u_i)(|u_i| - \lambda)_+ \tag{21}$$

Where $i \in [1, \ldots, K]$ describe the number of vectors that span the partial least squares subspace, calculated from the iterative

deflation: $i - 1$ of matrices Y and X. The iterative approach used was previously described in literature illustrating its application for sparse PCA (Shen and Huang, 2008), and the approximation of the LASSO regularisation function used was a soft thresholding approach, where only the positive values of u_i where used as the sparse vectors. This method has been included in the R package *mixOmics* (Rohart et al., 2017). Sparse PLS-DA was shown to improve classification accuracy on previously published datasets in chemometrics, although manipulation of the regularisation coefficient in addition to the number latent variables certainly adds a layer of complexity to the analysis (Filzmoser et al., 2012). It was also applied to distinguish between conventional and organic walnut oils by solid-phase micro-extraction GC-MS in a recent publication (Kalogiouri et al., 2021).

Also of note is a sparse discriminant analysis implemented using a Bayesian information criteria by Orlhac et al. (2019), and another sparse method using shrunken centroids (Chen et al., 2015)–both for simultaneous classification and feature selection.

In many of the previous citations, sparse methods for discriminant-type analyses are not used as a primary means of investigation, and it appears that for chemometrics research sparse methods are often included to the sake of comparison to existing methods. In some cases, it appears that sparse methods do not improve upon the classification accuracy versus more standard chemometrics tools such as PLS-DA, or reduction of the data dimensionality via PCA prior to an analysis of the resultant scores by CVA (Buzzini et al., 2021). Sparse methods may also suffer from the drawbacks associated with other embedded methods for feature selection, such as a tendency to over-fit to the training set. For hyphenated chromatographic-mass spectrometric data where analysts have come to expect a high level of missing features from their datasets, a feature selection method that optimises some criterion of the training set, may not correctly predict the test or validation set. This problem may be attributed to the high sensitivity of hyphenated instruments, but a complacency with low industry standards for data pre-processing is also suspected (Lu et al., 2008). It is also worth noting that most sparse methods for discriminant analysis or regression appear to be published in the R programming language, and since the working language for much of chemometrics is MATLAB, this could be another possible reason why these methods are not widely applied.

4.2 Sparse Projection Pursuit Analysis

Kurtosis minimisation as a projection index has long been utilised in chemometrics and performs well for classification problems, since scores with low measures of kurtosis typically are well-resolved within a linear subspace (Hou and Wentzell, 2011; Wentzell et al., 2021a,b). Kurtosis is described as the fourth statistical moment, following mean, variance, and skew (**Equation 22**). Distributions with a high degree of kurtosis describe data with higher tendency for outliers relative to a normal distribution. For the purposes of revealing clustering of the data, distributions with a low value of kurtosis can indicate a bimodality in their distribution. This bimodality naturally lends to clusters that are readily observable in higher dimensions, which are calculated step-wise for univariate measures of kurtosis (**Equation 23**), but can also be calculated simultaneously via a multivariate determination of kurtosis (**Eq. 24**) to avoid categorising samples into nominal clusters. What is interesting about PPA is that it is an unsupervised method, and does not calculate the subspace in a way that is informed by class information. It has been argued that this makes it less prone to over-fitting (Hou and Wentzell, 2011); however, it performs poorly for ill-posed problems and a dimensionality reduction step has historically been applied prior to its use. Being an unsupervised method, the use of a separate classifier is typically required following the analysis.

$$K = \frac{\frac{1}{m}\sum_{i=1}^{m}(z_i - \bar{z})^4}{\frac{1}{m}\left(\sum_{i=1}^{m}(z_i - \bar{z})^2\right)^2} \tag{22}$$

$$K = \frac{m\sum_{i=1}^{m}\left(\mathbf{v}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{v}\right)^2}{\left(\mathbf{v}^T X^T X\mathbf{v}\right)^2} \tag{23}$$

$$K = m\sum_{i=1}^{m}\left(tr\left(\left(V^T X^T XV\right)^{-1}\left(V^T\mathbf{x}_i\mathbf{x}_i^T V\right)\right)\right)^2 \tag{24}$$

In **Equation 22** the kurtosis of the i^{th} sample's projection (as z_i) to a latent variable space is described for all samples $i \in [1, m]$ where m is the total number of samples for an $m \times n$ matrix. Minimisation of **Eqs 23**, **24** as a function of the latent variables individually ($\mathbf{v}$) or collectively (V), operates on the sum of the projections of each sample as $\mathbf{x_i}$.

Hou and Wentzell (2014) implemented a regularisation parameter to the minimisation of kurtosis via a quasi-power algorithm. More recently, Driscoll et al. (2019) utilised a genetic algorithm to reduce the dimensionality of the feature set in order to perform PPA using kurtosis minimisation as a projection index (kPPA).

5 HYBRID, AND MISCELLANEOUS APPROACHES

5.1 Decision Trees

Decision trees, or more specifically Classification and Regression Trees (CART) are trees that perform binary operations based on some information criteria of the variable such as entropy or Gini impurity to establish a threshold for the decision (Questier et al., 2005). Each tree is comprised of nodes (τ) that are connected by branches. Each node is a point at which a decision is made about a particular sample given the information encoded by a single variable. If one node leads to two other nodes, it is described as a parent node–otherwise it is considered a terminal node, wherein a final decision regarding the class membership of a sample is made by weighing the outcomes of the decisions made previously further up the tree.

The heterogeneity, or disorder of each decision can be measured using entropy ($\sum_{k=0}^{1} - p_k log_2(p_k)$). Due to the logarithmic operator however, this term is computationally expensive. The Gini impurity, ($\Delta_k = \sum_{k=0}^{1} p_k(1 - p_k)$) is more efficient to calculate and closely resembles entropy. In either case, p_k is the probability of a given sample being labelled as the k^{th} class based on the characteristic being measured at a particular node.

Further details about decision trees can be found in an article by Zhang et al. (2005).

5.2 Random Forests

A Random Forest (RF) is comprised of a large number of decision trees, T that vote on the membership of an unknown sample. RFs are frequently used as a variable selection step, since the method largely evades the issue of dimensionality through a number of randomly determined variable subsets voting on the model outcome. RF is also a method that is especially simple to use, with the only critical parameter being the number of decision trees to include. This parameter is easy to optimise, since the number of decision trees usually scales with the relative dimensionality of the data. The functionality of RF can be used to overcome problems related to over-fitting of individual decision trees, through the consensus of a majority of parallel models. RFs are a widely-used tool in data analysis broadly speaking, and have long been used for Quantitative Structure-Activity Relationship (QSAR) modelling (Svetnik et al., 2003) in particular.

Random Forests (RFs) Breiman (2001) are difficult to classify as belonging to one of either filter, wrapper or embedded approaches. RFs are large collections of decision trees used to vote on the class membership of a sample based on its characteristics. The randomly generated variable subsets for each decision tree liken the method to wrapper methods, but since the method performs both classification and feature selection simultaneously Menze et al. (2009), they are also widely considered to fall under the umbrella of embedded methods. For interpretation of the most significant features; however, a threshold of variable significance is typically employed as part of the analysis, which could also qualify RFs as a filter method. For the purposes of this review, feature selection by RFs will be classified as a "hybrid" method, if only to signify that it is somewhat of an outlier compared with the methods that have been previously discussed.

Similar to PLS-DA and PPA, the importance of each variable in RF can be summarised by their relative importance in the model. In RF, the Gini importance is a measure of how often a variable was used to split the data across multiple decision trees, considering its discriminating value via For the j^{th} variable:

$$Gini(j) = \sum_{t=1}^{T} \sum_{r=1}^{\tau} \Delta_k(\tau, T) \tag{25}$$

Where Δ_k indicates the Gini impurity, or the ability of a variable to separate two classes (in a binary example), and $Gini(j)$ is the sum of this discriminatory power summed over all nodes (τ) and trees (T) in the RF model (Menze et al., 2009). The Gini coefficient has been used for feature selection routines for the detection of bovine spongiform encephalopathy via serum (Menze et al., 2007).

Mean decrease in accuracy (MDI) is also sometimes used for variable selection with RF, and describes the difference in prediction accuracy when a considered variable is excluded from the model. This method was used in conjunction with VIP scores using MetaboAnalyst by Azizan et al. (2021) to the end of detecting lard adulteration using fatty acids as analysed by GC-MS.

Although variable impact can be assessed using RFs, the exact manner in which they correlate with each other is not immediately clear using RFs. This is due to the fact that both decision trees and RFs are non-linear methods, and as such cannot be summarised by considering the co-linearity of variables directly. With that being said, it is nonetheless possible to interpret the selected features in the context of a linear model (Azizan et al., 2021). Despite its limited application in chemometrics for classification problems, RF has been included in a number of popular data analysis suites, including MetaboAnalyst (Pang et al., 2021) and ChromCompare+ (an analysis suite for GC×GC-TOFMS data), likely owing to the simplicity of its operation for inexperienced users.

5.3 Hybrid Methods

The cluster resolution function has been used in a number of studies as an objective function to select variables *via* a hybrid backwards elimination/forward selection approach (de la Mata et al., 2017; Nam et al., 2020; Sorochan Armstrong et al., 2022). The algorithm used for these studies, frequently called Feature Selection by Cluster Resolution (FS-CR), proceeds once the variables have been ranked, and appropriate cutoffs determined for backwards elimination to start, and for forward selection to end (as mentioned earlier, the start and stop numbers).

Hybridisation of wrapper and filter methods are common for assessing an initial subset of variables to perform backwards elimination or forward selection on, or for intelligently selecting a number of variables to consider using other subset selection methods (Zhang et al., 2019; Singh and Singh, 2021).

6 CONCLUSION

Analysts demand the most predictive subset of features using the fewest number of parameters possible. Fewer parameters typically rely more strongly on well-informed methodologies to optimise variable subsets, and offer fewer avenues of recourse should the routine fail to correctly indicated external samples to the model. Routines with more parameters to optimise may be more dependent on a skilled analyst, who may explore a number of avenues to gain better insight into the classification problem. A number of factors must be considered, including the relative dimensionality of the data and/or the total number of features (which may scale poorly regardless of the number of samples depending on what technique is being used), the number of missing elements in the data, the number and severity of outliers present in the dataset, and of course the instrumentation being used. The effect of outliers can be explored using unsupervised methods like PCA or PPA, and the effect of missing data can be observed by projecting previously unconsidered samples into the model to assess potential over-fitting.

The vast majority of feature selection routines that see frequent use in chemometrics return a subset of variables

whose linear combination can provide adequate discrimination, assuming that the instrument was operated within its linear range during the data acquisition step. However despite the fact that linear methods can account for co-linear variables, some data may require non-linear methods for subset determination or evaluation if the variables or combinations thereof are not sufficiently informative. Non-linear methods for variable selection and evaluation such as genetic algorithms or RFs can offer more discriminating power, but subsequent model interpretation may be difficult, and extensive user experimentation may be required for methods based on genetic algorithms to optimise a high number of user-input parameters.

A recent study by Vrábel et al. (2020) examined the result of a contest for a challenging classification problem based on laser-induced breakdown spectroscopy. Each team used various combinations of either linear or non-linear approaches for feature selection and classification, but the winning team was the one that focused on manual data exploration and interpretation to inform a simple classifier using PLS-DA. This study highlights the conventional wisdom that human interpretation and insight is difficult to beat using machines, regardless of the technique used.

The orthodoxy of linear modelling and feature selection in chemometrics may yet be challenged at some point in the future, where feature selection tools based on highly non-linear methods (Upadhyay et al., 2020; Ranjan et al., 2021) are eventually explored using chemical data. Uniform Manifold Approximation and Projection (UMAP) is generating a lot of interest in fields adjacent to chemometrics, in particular the -omics fields, where the underlying biological phenomena may not always be described using the same linear assumptions that are suitable for most chemical analyses (Shen et al., 2020).

AUTHOR CONTRIBUTIONS

MSA performed all research, and summarisation for the article. AdlM reviewed the article for correctness. JH was responsible for conceptualisation, and also reviewed the article for correctness.

FUNDING

The authors acknowledge the support provided by the Natural Sciences and Engineering Council of Canada (NSERC), and funding provided by Genome Canada, Genome Alberta, and the Canada Foundation for Innovation that support the Metabolomics Innovation Centre (TMIC).

REFERENCES

Adutwum, L. A., de la Mata, A. P., Bean, H. D., Hill, J. E., and Harynuk, J. J. (2017). Estimation of Start and Stop Numbers for Cluster Resolution Feature Selection Algorithm: an Empirical Approach Using Null Distribution Analysis of Fisher Ratios. *Anal. Bioanal. Chem.* 409, 6699–6708. doi:10.1007/s00216-017-0628-8

Aliakbarzadeh, G., Parastar, H., and Sereshti, H. (2016). Classification of Gas Chromatographic Fingerprints of Saffron Using Partial Least Squares Discriminant Analysis Together with Different Variable Selection Methods. *Chemom. Intelligent Laboratory Syst.* 158, 165–173. doi:10.1016/j.chemolab.2016.09.002

Amante, E., Salomone, A., Alladio, E., Vincenti, M., Porpiglia, F., and Bro, R. (2019). Untargeted Metabolomic Profile for the Detection of Prostate Carcinoma-Preliminary Results from PARAFAC2 and PLS-DA Models. *Molecules* 24, 3063. doi:10.3390/molecules24173063

Andersen, C. M., and Bro, R. (2010). Variable Selection in Regression-A Tutorial. *J. Chemom.* 24, 728–737. doi:10.1002/cem.1360

Armstrong, M. S., de la Mata, A. P., and Harynuk, J. J. (2021). An Efficient and Accurate Numerical Determination of the Cluster Resolution Metric in Two Dimensions. *J. Chemom.* 35 (7-8), e3346. doi:10.1002/cem.3346

Azizan, N. I., Mokhtar, N. F. K., Arshad, S., Sharin, S. N., Mohamad, N., Mustafa, S., et al. (2021). Detection of Lard Adulteration in Wheat Biscuits Using Chemometrics-Assisted Gcms and Random Forest. *Food Anal. Methods* 14, 1–12. doi:10.1007/s12161-021-02046-9

Ballabio, D., Skov, T., Leardi, R., and Bro, R. (2008). Classification of Gc-Ms Measurements of Wines by Combining Data Dimension Reduction and Variable Selection Techniques. *J. Chemom.* 22, 457–463. doi:10.1002/cem.1173

Box, G. E. P. (1953). Non-Normality and Tests on Variances. *Biometrika* 40, 318–335. doi:10.1093/biomet/40.3-4.318

Breiman, L. (2001). Random Forests. *Mach. Learn.* 45, 5–32. doi:10.1023/a:1010933404324

Buzzini, P., Curran, J., and Polston, C. (2021). Comparison between Visual Assessments and Different Variants of Linear Discriminant Analysis to the Classification of Raman Patterns of Inkjet Printer Inks. *Forensic Chem.* 24, 100336. doi:10.1016/j.forc.2021.100336

Centner, V., Massart, D.-L., de Noord, O. E., de Jong, S., Vandeginste, B. M., and Sterna, C. (1996). Elimination of Uninformative Variables for Multivariate Calibration. *Anal. Chem.* 68, 3851–3858. doi:10.1021/ac960321m

Chen, C., Zhang, Z.-M., Ouyang, M.-L., Liu, X., Yi, L., and Liang, Y.-Z. (2015). Shrunken Centroids Regularized Discriminant Analysis as a Promising Strategy for Metabolomics Data Exploration. *J. Chemom.* 29, 154–164. doi:10.1002/cem.2685

Chong, I.-G., and Jun, C.-H. (2005). Performance of Some Variable Selection Methods when Multicollinearity Is Present. *Chemom. intelligent laboratory Syst.* 78, 103–112. doi:10.1016/j.chemolab.2004.12.011

Chun, H., and Keleş, S. (2010). Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 72, 3–25. doi:10.1111/j.1467-9868.2009.00723.x

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011). Sparse Discriminant Analysis. *Technometrics* 53, 406–413. doi:10.1198/tech.2011.08118

Clemmensen, L., and Kuhn, M. M. (2016). *Package 'sparselda'*. CRAN Repository.

Cocchi, M., Biancolillo, A., and Marini, F. (2018). "Data Analysis for Omic Sciences: Methods and Applicationsof Comprehensive Analytical Chemistry," in *Chap. Chemometric Methods for Classification and Feature Selection*. Editors J. Jaumot, and R. Tauler (Elsevier), 82, 265–299. doi:10.1016/bs.coac.2018.08.006

Coon, A. M., Beyramysoltan, S., and Musah, R. A. (2019). A Chemometric Strategy for Forensic Analysis of Condom Residues: Identification and Marker Profiling of Condom Brands from Direct Analysis in Real Time-High Resolution Mass Spectrometric Chemical Signatures. *Talanta* 194, 563–575. doi:10.1016/j.talanta.2018.09.101

Crammer, K., and Singer, Y. (2001). On the Algorithmic Implementation of Multiclass Kernel-Based Vector Machines. *J. Mach. Learn. Res.* 2, 265–292.

de Andrade, B. M., de Gois, J. S., Xavier, V. L., and Luna, A. S. (2020). Comparison of the Performance of Multiclass Classifiers in Chemical Data: Addressing the Problem of Overfitting with the Permutation Test. *Chemom. Intelligent Laboratory Syst.* 201, 104013. doi:10.1016/j.chemolab.2020.104013

de la Mata, A. P., McQueen, R. H., Nam, S. L., and Harynuk, J. J. (2017). Comprehensive Two-Dimensional Gas Chromatographic Profiling and Chemometric Interpretation of the Volatile Profiles of Sweat in Knit Fabrics. *Anal. Bioanal. Chem.* 409, 1905–1913. doi:10.1007/s00216-016-0137-1

Dettmer, K., Aronov, P. A., and Hammock, B. D. (2007). Mass Spectrometry-Based Metabolomics. *Mass Spectrom. Rev.* 26, 51–78. doi:10.1002/mas.20108

Driscoll, S. P., MacMillan, Y. S., and Wentzell, P. D. (2019). Sparse Projection Pursuit Analysis: an Alternative for Exploring Multivariate Chemical Data. *Anal. Chem.* 92, 1755–1762. doi:10.1021/acs.analchem.9b03166

Farrés, M., Platikanov, S., Tsakovski, S., and Tauler, R. (2015). Comparison of the Variable Importance in Projection (Vip) and of the Selectivity Ratio (Sr) Methods for Variable Selection and Interpretation. *J. Chemom.* 29, 528–536. doi:10.1002/cem.2736

Filzmoser, P., Gschwandtner, M., and Todorov, V. (2012). Review of Sparse Methods in Regression and Classification with Application to Chemometrics. *J. Chemom.* 26, 42–51. doi:10.1002/cem.1418

Ghisoni, S., Lucini, L., Rocchetti, G., Chiodelli, G., Farinelli, D., Tombesi, S., et al. (2020). Untargeted Metabolomics with Multivariate Analysis to Discriminate Hazelnut (Corylus Avellana l.) Cultivars and Their Geographical Origin. *J. Sci. Food Agric.* 100, 500–508. doi:10.1002/jsfa.9998

Hopke, P. K. (2003). The Evolution of Chemometrics. *Anal. Chim. Acta* 500, 365–377. doi:10.1016/s0003-2670(03)00944-9

Hou, S., and Wentzell, P. D. (2014). Regularized Projection Pursuit for Data with a Small Sample-To-Variable Ratio. *Metabolomics* 10, 589–606. doi:10.1007/s11306-013-0612-z

Hou, S., and Wentzell, P. (2011). Fast and Simple Methods for the Optimization of Kurtosis Used as a Projection Pursuit Index. *Anal. Chim. acta* 704, 1–15. doi:10.1016/j.aca.2011.08.006

Johnson, K. J., and Synovec, R. E. (2002). Pattern Recognition of Jet Fuels: Comprehensive Gc× Gc with Anova-Based Feature Selection and Principal Component Analysis. *Chemom. Intelligent Laboratory Syst.* 60, 225–237. doi:10.1016/s0169-7439(01)00198-8

Kalogiouri, N. P., Manousi, N., Rosenberg, E., Zachariadis, G. A., Paraskevopoulou, A., and Samanidou, V. (2021). Exploring the Volatile Metabolome of Conventional and Organic Walnut Oils by Solid-phase Microextraction and Analysis by Gc-Ms Combined with Chemometrics. *Food Chem.* 363, 130331. doi:10.1016/j.foodchem.2021.130331

Kohavi, R., and John, G. H. (1997). Wrappers for Feature Subset Selection. *Artif. Intell.* 97, 273–324. doi:10.1016/s0004-3702(97)00043-x

Kvalheim, O. M. (2020). Variable Importance: Comparison of Selectivity Ratio and Significance Multivariate Correlation for Interpretation of Latent-Variable Regression Models. *J. Chemom.* 34, e3211. doi:10.1002/cem.3211

Lavine, B. K., Brzozowski, D., Moores, A. J., Davidson, C., and Mayfield, H. T. (2001). Genetic Algorithm for Fuel Spill Identification. *Anal. Chim. Acta* 437, 233–246. doi:10.1016/s0003-2670(01)00946-1

Lavine, B. K., Nuguru, K., and Mirjankar, N. (2011). One Stop Shopping: Feature Selection, Classification and Prediction in a Single Step. *J. Chemom.* 25, 116–129. doi:10.1002/cem.1358

Lavine, B. K., Ritter, J., Moores, A. J., Wilson, M., Faruque, A., and Mayfield, H. T. (2000). Source Identification of Underground Fuel Spills by Solid-phase Microextraction/high-Resolution Gas Chromatography/genetic Algorithms. *Anal. Chem.* 72, 423–431. doi:10.1021/ac9904967

Lê Cao, K.-A., Boitard, S., and Besse, P. (2011). Sparse Pls Discriminant Analysis: Biologically Relevant Feature Selection and Graphical Displays for Multiclass Problems. *BMC Bioinforma.* 12, 1–17. doi:10.1186/1471-2105-12-253

Lu, H., Liang, Y., Dunn, W. B., Shen, H., and Kell, D. B. (2008). Comparative Evaluation of Software for Deconvolution of Metabolomics Data Based on Gc-Tof-Ms. *TrAC Trends Anal. Chem.* 27, 215–227. doi:10.1016/j.trac.2007.11.004

Maddala, G. S., and Lahiri, K. (1992). *Introduction to Econometrics*, 2. New York: Macmillan.

Marney, L. C., Siegler, W. C., Parsons, B. A., Hoggard, J. C., Wright, B. W., and Synovec, R. E. (2013). Tile-based Fisher-Ratio Software for Improved Feature Selection Analysis of Comprehensive Two-Dimensional Gas Chromatography–Time-Of-Flight Mass Spectrometry Data. *Talanta* 115, 887–895. doi:10.1016/j.talanta.2013.06.038

Mehmood, T., Liland, K. H., Snipen, L., and Sæbø, S. (2012). A Review of Variable Selection Methods in Partial Least Squares Regression. *Chemom. intelligent laboratory Syst.* 118, 62–69. doi:10.1016/j.chemolab.2012.07.010

Mehmood, T., Sæbø, S., and Liland, K. H. (2020). Comparison of Variable Selection Methods in Partial Least Squares Regression. *J. Chemom.* 34, e3226. doi:10.1002/cem.3226

Menze, B. H., Kelm, B. M., Masuch, R., Himmelreich, U., Bachert, P., Petrich, W., et al. (2009). A Comparison of Random Forest and its Gini Importance with Standard Chemometric Methods for the Feature Selection and Classification of Spectral Data. *BMC Bioinforma.* 10, 1–16. doi:10.1186/1471-2105-10-213

Menze, B. H., Petrich, W., and Hamprecht, F. A. (2007). Multivariate Feature Selection and Hierarchical Classification for Infrared Spectroscopy: Serum-Based Detection of Bovine Spongiform Encephalopathy. *Anal. Bioanal. Chem.* 387, 1801–1807. doi:10.1007/s00216-006-1070-5

Nam, S. L., de la Mata, A. P., Dias, R. P., and Harynuk, J. J. (2020). Towards Standardization of Data Normalization Strategies to Improve Urinary Metabolomics Studies by Gc× Gc-Tofms. *Metabolites* 10, 376. doi:10.3390/metabo10090376

Nørgaard, L., Bro, R., Westad, F., and Engelsen, S. B. (2006). A Modification of Canonical Variates Analysis to Handle Highly Collinear Multivariate Data. *J. Chemom. A J. Chemom. Soc.* 20, 425–435.

Nørgaard, L., Saudland, A., Wagner, J., Nielsen, J. P., Munck, L., and Engelsen, S. B. (2000). Interval Partial Least-Squares Regression (I Pls): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy. *Appl. Spectrosc.* 54, 413–419.

Orlhac, F., Mattei, P.-A., Bouveyron, C., and Ayache, N. (2019). Class-specific Variable Selection in High-Dimensional Discriminant Analysis through Bayesian Sparsity. *J. Chemom.* 33, e3097. doi:10.1002/cem.3097

Pang, Z., Chong, J., Zhou, G., de Lima Morais, D. A., Chang, L., Barrette, M., et al. (2021). Metaboanalyst 5.0: Narrowing the Gap between Raw Spectra and Functional Insights. *Nucleic acids Res.* 49 (W1), W388–W396. doi:10.1093/nar/gkab382

Peris-Díaz, M. D., Łydżba-Kopczyńska, B., and Sentandreu, E. (2018). Raman Spectroscopy Coupled to Chemometrics to Discriminate Provenance and Geological Age of Amber. *J. Raman Spectrosc.* 49, 842–851. doi:10.1002/jrs.5357

Pesesse, R., Stefanuto, P.-H., Schleich, F., Louis, R., and Focant, J.-F. (2019). Multimodal Chemometric Approach for the Analysis of Human Exhaled Breath in Lung Cancer Patients by Td-Gc× Gc-Tofms. *J. Chromatogr. B* 1114, 146–153. doi:10.1016/j.jchromb.2019.01.029

Pierce, K. M., Hoggard, J. C., Hope, J. L., Rainey, P. M., Hoofnagle, A. N., Jack, R. M., et al. (2006). Fisher Ratio Method Applied to Third-Order Separation Data to Identify Significant Chemical Components of Metabolite Extracts. *Anal. Chem.* 78, 5068–5075. doi:10.1021/ac0602625

Questier, F., Put, R., Coomans, D., Walczak, B., and Vander Heyden, Y. (2005). The Use of Cart and Multivariate Regression Trees for Supervised and Unsupervised Feature Selection. *Chemom. Intelligent Laboratory Syst.* 76, 45–54. doi:10.1016/j.chemolab.2004.09.003

Rajalahti, T., Arneberg, R., Berven, F. S., Myhr, K.-M., Ulvik, R. J., and Kvalheim, O. M. (2009). Biomarker Discovery in Mass Spectral Profiles by Means of Selectivity Ratio Plot. *Chemom. Intelligent Laboratory Syst.* 95, 35–48. doi:10.1016/j.chemolab.2008.08.004

Ranjan, B., Sun, W., Park, J., Mishra, K., Schmidt, F., Xie, R., et al. (2021). Dubstepr Is a Scalable Correlation-Based Feature Selection Method for Accurately Clustering Single-Cell Data. *Nat. Commun.* 12, 1–12. doi:10.1038/s41467-021-26085-2

Rich, D. C., Livingston, K. M., and Morgan, S. L. (2020). Evaluating Performance of Lasso Relative to Pca and Lda to Classify Dyes on Fibers. *Forensic Chem.* 18, 100213. doi:10.1016/j.forc.2020.100213

Rinnan, Å., Andersson, M., Ridder, C., and Engelsen, S. B. (2014). Recursive Weighted Partial Least Squares (Rpls): an Efficient Variable Selection Method Using Pls. *J. Chemom.* 28, 439–447. doi:10.1002/cem.2582

Rohart, F., Gautier, B., Singh, A., and Lê Cao, K.-A. (2017). Mixomics: An R Package for 'omics Feature Selection and Multiple Data Integration. *PLoS Comput. Biol.* 13, e1005752. doi:10.1371/journal.pcbi.1005752

Santosa, F., and Symes, W. W. (1986). Linear Inversion of Band-Limited Reflection Seismograms. *SIAM J. Sci. Stat. Comput.* 7, 1307–1330. doi:10.1137/0907087

Sorochan Armstrong, M. D., Arredondo Campos, O. R., Bannon, C. C., de la Mata, A. P., Case, R. J., and Harynuk, J. J. (2022). Global Metabolome Analysis of Dunaliella Tertiolecta, Phaeobacter Italicus R11 Co-cultures Using Thermal Desorption - Comprehensive Two-Dimensional Gas Chromatography - Time-Of-Flight Mass Spectrometry (TD-GC×GC-TOFMS). *Phytochemistry* 195, 113052. doi:10.1016/j.phytochem.2021.113052

Seijo, S., Lozano, J. J., Alonso, C., Reverter, E., Miquel, R., Abraldes, J. G., et al. (2013). Metabolomics Discloses Potential Biomarkers for the Noninvasive

Diagnosis of Idiopathic Portal Hypertension. *Official J. Am. Coll. Gastroenterology— ACG* 108, 926–932. doi:10.1038/ajg.2013.11

Shen, B., Yi, X., Sun, Y., Bi, X., Du, J., Zhang, C., et al. (2020). Proteomic and Metabolomic Characterization of Covid-19 Patient Sera. *Cell* 182, 59–72. doi:10.1016/j.cell.2020.05.032

Shen, H., and Huang, J. Z. (2008). Sparse Principal Component Analysis via Regularized Low Rank Matrix Approximation. *J. Multivar. analysis* 99, 1015–1034. doi:10.1016/j.jmva.2007.06.007

Sinclair, E., Trivedi, D. K., Sarkar, D., Walton-Doyle, C., Milne, J., Kunath, T., et al. (2021). Metabolomics of Sebum Reveals Lipid Dysregulation in Parkinson's Disease. *Nat. Commun.* 12, 1–9. doi:10.1038/s41467-021-21669-4

Singh, N., and Singh, P. (2021). A Hybrid Ensemble-Filter Wrapper Feature Selection Approach for Medical Data Classification. *Chemom. Intelligent Laboratory Syst.* 217, 104396. doi:10.1016/j.chemolab.2021.104396

Sinkov, N. A., and Harynuk, J. J. (2011). Cluster Resolution: A Metric for Automated, Objective and Optimized Feature Selection in Chemometric Modeling. *Talanta* 83, 1079–1087. doi:10.1016/j.talanta.2010.10.025

Sinkov, N. A., Johnston, B. M., Sandercock, P. M. L., and Harynuk, J. J. (2011). Automated Optimization and Construction of Chemometric Models Based on Highly Variable Raw Chromatographic Data. *Anal. Chim. acta* 697, 8–15. doi:10.1016/j.aca.2011.04.029

Stoessel, D., Stellmann, J.-P., Willing, A., Behrens, B., Rosenkranz, S. C., Hodecker, S. C., et al. (2018). Metabolomic Profiles for Primary Progressive Multiple Sclerosis Stratification and Disease Course Monitoring. *Front. Hum. Neurosci.* 12, 226. doi:10.3389/fnhum.2018.00226

Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., and Feuston, B. P. (2003). Random Forest: a Classification and Regression Tool for Compound Classification and Qsar Modeling. *J. Chem. Inf. Comput. Sci.* 43, 1947–1958. doi:10.1021/ci034160g

Talukdar, U., Hazarika, S. M., and Gan, J. Q. (2018). A Kernel Partial Least Square Based Feature Selection Method. *Pattern Recognit.* 83, 91–106. doi:10.1016/j.patcog.2018.05.012

Theodoridis, S. (2020). "Chapter 7 - Classification: a Tour of the Classics," in *Machine Learning*. Editor S. Theodoridis. Second EditionSecond edition edn (Academic Press), 301–350. doi:10.1016/B978-0-12-818803-3.00016-7

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288. doi:10.1111/j.2517-6161.1996.tb02080.x

Tibshirani, R. J. (2013). The Lasso Problem and Uniqueness. *Electron. J. statistics* 7, 1456–1490. doi:10.1214/13-ejs815

Tran, T. N., Afanador, N. L., Buydens, L. M., and Blanchet, L. (2014). Interpretation of Variable Importance in Partial Least Squares with Significance Multivariate Correlation (Smc). *Chemom. Intelligent Laboratory Syst.* 138, 153–160. doi:10.1016/j.chemolab.2014.08.005

Trendafilov, N. T., and Jolliffe, I. T. (2007). Dalass: Variable Selection in Discriminant Analysis via the Lasso. *Comput. Statistics Data Analysis* 51, 3718–3736. doi:10.1016/j.csda.2006.12.046

Upadhyay, D., Manero, J., Zaman, M., and Sampalli, S. (2020). Gradient Boosting Feature Selection with Machine Learning Classifiers for Intrusion Detection on Power Grids. *IEEE Trans. Netw. Serv. Manag.* 18, 1104–1116.

Van der Maaten, L., and Hinton, G. (2008). Visualizing Data Using T-Sne. *J. Mach. Learn. Res.* 9, 2579–2605.

Vrábel, J., Képeš, E., Duponchel, L., Motto-Ros, V., Fabre, C., Connemann, S., et al. (2020). Classification of Challenging Laser-Induced Breakdown Spectroscopy Soil Sample Data-Emslibs Contest. *Spectrochim. Acta Part B At. Spectrosc.* 169, 105872. doi:10.1016/j.sab.2020.105872

Wang, R., and Tang, K. (2009). "Feature Selection for Maximizing the Area under the Roc Curve," in 2009 IEEE International Conference on Data Mining Workshops (IEEE), 400–405. doi:10.1109/icdmw.2009.25

Wentzell, P. D., Giglio, C., and Kompany-Zareh, M. (2021a). Beyond Principal Components: a Critical Comparison of Factor Analysis Methods for Subspace Modelling in Chemistry. *Anal. Methods* 13, 4188–4219. doi:10.1039/d1ay01124c

Wentzell, P. D., Gonçalves, T. R., Matsushita, M., and Valderrama, P. (2021b). Combinatorial Projection Pursuit Analysis for Exploring Multivariate Chemical Data. *Anal. Chim. Acta*, 338716. doi:10.1016/j.aca.2021.338716

Witten, D. M., and Tibshirani, R. (2011). Penalized Classification Using Fisher's Linear Discriminant. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 73, 753–772. doi:10.1111/j.1467-9868.2011.00783.x

Witten, D., and Witten, M. D. (2015). Package 'penalizedlda'. Package Penalized Classification Using Fishers's Linear Discriminant. Available at: https://cran.r-project.org/web/packages/penalizedLDA/penalizedLDA.pdf (accessed April 19, 2021).

Wold, S., Johansson, E., and Cocchi, M. (1993). *3d Qsar in Drug Design: Theory, Methods and Applications*. Leiden, Holland: ESCOM, 523–550.

Wold, S., Sjöström, M., and Eriksson, L. (2001). Pls-regression: a Basic Tool of Chemometrics. *Chemom. intelligent laboratory Syst.* 58, 109–130. doi:10.1016/s0169-7439(01)00155-1

Yang, S., Li, C., Mei, Y., Liu, W., Liu, R., Chen, W., et al. (2021). Discrimination of Corn Variety Using Terahertz Spectroscopy Combined with Chemometrics Methods. *Spectrochimica Acta Part A Mol. Biomol. Spectrosc.* 252, 119475. doi:10.1016/j.saa.2021.119475

Yendle, P. W., and MacFie, H. J. (1989). Discriminant Principal Components Analysis. *J. Chemom.* 3, 589–600. doi:10.1002/cem.1180030407

Zhang, J., Xiong, Y., and Min, S. (2019). A New Hybrid Filter/wrapper Algorithm for Feature Selection in Classification. *Anal. Chim. acta* 1080, 43–54. doi:10.1016/j.aca.2019.06.054

Zhang, M., Xu, Q., Daeyaert, F., Lewi, P., and Massart, D. (2005). Application of Boosting to Classification Problems in Chemometrics. *Anal. Chim. Acta* 544, 167–176. doi:10.1016/j.aca.2005.01.075

Zhu, Y., and Tan, T. L. (2016). Penalized Discriminant Analysis for the Detection of Wild-Grown and Cultivated Ganoderma Lucidum Using Fourier Transform Infrared Spectroscopy. *Spectrochimica Acta Part A Mol. Biomol. Spectrosc.* 159, 68–77. doi:10.1016/j.saa.2016.01.018

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.