Analyst

CRITICAL REVIEW

View Article Online
View Journal | View Issue

Cite this: *Analyst*, 2018, **143**, 3526

Partial least squares-discriminant analysis (PLS-DA) for classification of high-dimensional (HD) data: a review of contemporary practice strategies and knowledge gaps†

Loong Chuen Lee, *[a,b] Choong-Yeun Liong *[b] and Abdul Aziz Jemain [b]

Partial least squares-discriminant analysis (PLS-DA) is a versatile algorithm that can be used for predictive and descriptive modelling as well as for discriminative variable selection. However, versatility is both a blessing and a curse and the user needs to optimize a wealth of parameters before reaching reliable and valid outcomes. Over the past two decades, PLS-DA has demonstrated great success in modelling high-dimensional datasets for diverse purposes, *e.g.* product authentication in food analysis, diseases classification in medical diagnosis, and evidence analysis in forensic science. Despite that, in practice, many users have yet to grasp the essence of constructing a valid and reliable PLS-DA model. As the technology progresses, across every discipline, datasets are evolving into a more complex form, *i.e.* multi-class, imbalanced and colossal. Indeed, the community is welcoming a new era called big data. In this context, the aim of the article is two-fold: (a) to review, outline and describe the contemporary PLS-DA modelling practice strategies, and (b) to critically discuss the respective knowledge gaps that have emerged in response to the present big data era. This work could complement other available reviews or tutorials on PLS-DA, to provide a timely and user-friendly guide to researchers, especially those working in applied research.

Received 30th March 2018,
Accepted 31st May 2018

DOI: 10.1039/c8an00599k

rsc.li/analyst

1 Introduction

The partial least squares (PLS) algorithm was first introduced for regression tasks and then evolved into a classification method that is well known as PLS-discriminant analysis (PLS-DA).[1–5] In practice, the PLS-DA algorithm has been used for predictive and descriptive modelling, as well as discriminative variable selection. Herein, we restrict our attention primarily to the predictive modelling. Theoretically, PLS-DA combines dimensionality reduction and discriminant analysis into one algorithm and is especially applicable to modelling high-dimensional (HD) data. In addition, PLS-DA does not assume the data to fit a particular distribution and thus is more flexible than other discriminant algorithms like Fisher's linear discriminant analysis (LDA). Consequently, PLS-DA modelling has a myriad of applications that span diverse fields: forensic science, banking, medical diagnosis, food analysis, metabolomics and soil science.[6–13] In fact, one formal description of the algorithm published in 2003 has been cited over 1700 times based on Google Scholar since then.[14] Despite that, in reality, many users have yet to grasp the essence of constructing a valid and reliable PLS-DA model.

Mathematically, PLS-DA modelling is not a one-step procedure but involves a series of mathematical operations and a wealth of parameters. It is the first author's experience who is enthusiastic about the potential of PLS-DA in modelling infrared (IR) spectra for solving forensic-based problems,[15–17] but finds no work addressing systematically the general PLS-DA modelling practice strategies. Although some papers have described and critically discussed the pitfalls of PLS-DA,[2,6,18–21] we noticed that the decision rule (DR) and empirical differences between PLS1-DA *versus* PLS2-DA algorithms have not been elaborated in detail but only briefly discussed on the theoretical ground. In practice, a majority of the users (especially those without an in-depth knowledge of statistics) seldom provide sufficient details of the two aspects in their writing.[2] On the other hand, the intimate collaborations between engineering, computer science and analytical science have sped up the development of cutting edge analytical instruments. Consequently, across every discipline, the resulting datasets are getting bigger and more complex, *i.e.* multi-

[a]*Forensic Science Programme, FSK, Universiti Kebangsaan Malaysia, Jalan Raja Muda Abdul Aziz, 50300 Kuala Lumpur, Malaysia. E-mail: lc_lee@ukm.edu.my*
[b]*Statistics Programme, FST, Universiti Kebangsaan Malaysia, 43600 Bangi, Selangor, Malaysia. E-mail: lg@ukm.edu.my*

†Electronic supplementary information (ESI) available. See DOI: 10.1039/c8an00599k

class, imbalanced and colossal. In fact, a new era called big data is emerging in the literature.

In this context, the aim of the article is two-fold: (a) to outline and describe the contemporary PLS-DA modelling practice strategies; and (b) to critically discuss the respective knowledge gaps that have emerged in response to the present big data era. Our work is indeed complementing other similar studies, *e.g.* ref. 18–21, and can be used as a user-friendly guide to researchers, especially those working in applied research. The remainder of the paper is organized as follows: section 2 presents a brief explanation of the PLS-DA algorithm. Then, section 3 describes the contemporary PLS-DA modelling practice strategies. The respective knowledge gaps are also critically discussed in the same section. A brief conclusion is presented in section 4.

2 Theoretical background

Although detailed descriptions of the PLS-DA algorithm are numerous,[2–5] it is worth briefly summarizing the essence of the algorithm, especially for junior researchers who might not have learnt about PLS-DA before reading this paper. Fundamentally, the PLS-DA predictive modelling encompasses two main procedures: (a) PLS component construction (*i.e.* dimension reduction), and (b) prediction model construction (*i.e.* discriminant analysis).

Historically, PLS was proposed to handle continuous variables (*i.e.* regression task). But in classification task, the output variables will always be categorical. As such, the first step in PLS-DA modelling is recoding the categorical variables (*i.e.* ordinal or nominal) into continuous variables (*i.e.* numerical). Examples are datasets of IR spectra of two (binary, $G = 2$) and three (multi-class, $G > 2$) different pen sources, *e.g.* PILOT, STABILO and ZEBRA. The recoding of the categorical variables (*i.e.* pen names) into continuous variables (*i.e.* dummy codes +1 and 0) is as illustrated in Fig. 1. In a binary classification problem, **y** is recoded to consist of only two integers. Typically 0 and +1 are used to indicate 'out-group' and 'in-group', respectively. Sometimes −1 is used to denote 'out-group'. On the other hand, a multi-class problem would have converted **y** into a dummy **Y**. In practice, PLS1-DA and PLS2-DA algorithms respectively employ the dummy **y** and **Y** as output variables. A more concrete description of the two algorithms is given in section 3.

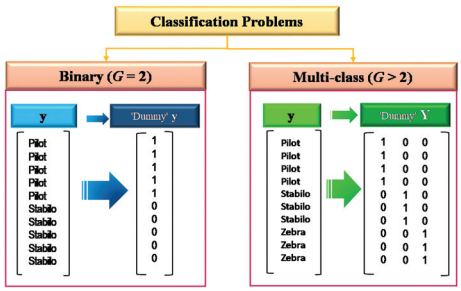

Fig. 1 Recoding of **y** into 'dummy' **y** or **Y**, respectively for binary (left) or multi-class (right) problem. The former is required in PLS1-DA and the latter is used in PLS2-DA modelling. **y**: categorical output variables; *G*: total number of classes; 1: 'in-group'; 0: 'out-group'.

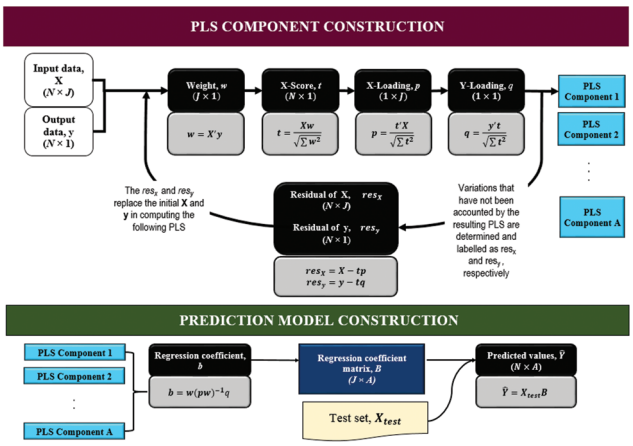

Fig. 2 Schematic flow of computational steps in estimating the *A* PLS components using PLS1-DA algorithm. *N*: number of samples, *J*: number of variables; **X**: input data; **y**: output data; $\boldsymbol{x}_{\text{test}}$: test sample.

Fig. 2 illustrates the procedures employed in constructing the *A* PLS component (*i.e.* new axes) *via* the PLS1-DA algorithm by using the input (**X**) and output (**y**) data.[2] It is worth noting here that the PLS2-DA algorithm would have employed dummy **Y** ($N \times G$) and not dummy **y** ($N \times 1$) in the first step. Firstly, the weight vector ($\boldsymbol{w}$) is estimated by maximizing the covariance between **X** and **y**. Following that, X-scores ($\boldsymbol{t}$), X-loading ($\boldsymbol{p}$) and Y-loading (q) are determined sequentially. Last but not least, regression coefficient ($\boldsymbol{b}$) is estimated using the resulting $\boldsymbol{w}$, $\boldsymbol{p}$ and q. Following that, the first set of PLS components and loading is established. Then the residuals **X** ($\text{res}_{\boldsymbol{X}}$) and **y** ($\text{res}_{\boldsymbol{y}}$) of the first PLS component become the input data (**X**) and output data (**y**), respectively, for constructing the second PLS component. The procedures are repeated *A* times if *A* number of PLS components are required to construct the desired prediction model. It is worth noting that in this algorithm, normalisation of the entire weight term is done in step 2. Eventually, *A* PLS components are constructed using the training samples, and a regression coefficient matrix, $\boldsymbol{B}$, is also prepared for subsequent prediction purpose.

For prediction purposes, the test set ($\boldsymbol{X}_{\text{test}}$) is reduced into the new dimensions (*i.e. A* PLS components) *via* **B** to produce the predicted values ($\hat{y}$), *i.e.* ypred. For the sake of brevity, ypred will be used consistently throughout the following discussions to denote the predicted values. Given a set of training data that belonged to *G* classes, the PLS-DA model would have produced *G* predicted values ($\hat{y} = \{y_1, \ldots, y_G\}$) for each test sample ($\boldsymbol{x}_{\text{test}}$). Table 1 presents the brief interpretations of the PLS-DA outcomes as listed in Fig. 2.

As has been described earlier, the perfect class membership, *i.e.* the predicted value ($\hat{y}$), is supposed to be ' + ' or '0' to indicate 'in-group' or 'out-group'. However, in practice, the

Table 1 Interpretations of the outcomes produced after the construction of A PLS components as illustrated in Fig. 2

Outcomes (dimensions)	Interpretations
Weight ($J \times A$)	Capture the maximum covariance of input (**X**) and output (**y**) data
X-scores ($N \times A$)	Coordinates of N training samples in A new axes
X-loadings ($A \times J$)	Coefficient between (J) raw input variables (**X**) and A PLS components
Y-loadings ($A \times 1$)	Coefficient between (1) raw dummy variables of y (in-class) and A PLS components
Regression coefficient ($J \times A$)	Contribution of (J) raw input variables (**X**) in A PLS components
Predicted values ($N \times A$)	Predicted class membership for N test samples in A new axes.

resulting predicted values often take on any values between 0 and 1 instead of an integer. For this reason, various decision rules (DRs) have been proposed in the literature to translate the predicted value into meaningful class membership. More details about the DR are reported in section 3.

3 Status quo in PLS-DA classification modelling practice strategies

3.1 Introduction

As emphasized in the previous section, PLS-DA is powerful in the sense that it produces multiple outcomes for predictive and descriptive modelling, and also variable selection. However, like a two-edged sword, researchers need to optimize a number of parameters before reaching an optimum model. In spite of its wide use and its generally good performances, several recent papers have been devoted to the pitfalls and misunderstandings in PLS-DA modeling.[2,18–21] Often, the users seldom consider each step carefully prior to choosing the parameter but tend to use the default option that is pre-selected in the statistical software.[2] Agreeing with this concern, we examined a number of studies published recently that have employed PLS-DA in order to gain insight into the contemporary practice strategies.

A total of 68 articles published since 2013 were reviewed[22–89] and the following information was recorded: (a) the nature and dimension of the dataset, (b) algorithm/acronym, (c) input data (*i.e.* global or interval selection), (d) data pre-processing (DP) methods, (e) model validation; (f) decision rules (DRs), (g) approach to determine the optimum number of PLS components, and (h) figures of merits employed to describe the model performance. The articles were chosen in such a way that the studies must have included PLS-DA for building a prediction model. Studies that employed PLS-DA only for exploratory purpose have not been included since this review is aiming to elucidate practice strategies in constructing a reliable PLS-DA prediction model.

The details of the survey results are reported in ESI Table 1.† The corresponding graphical summary is presented in Fig. 3. There are at least eight important practical aspects involved in the PLS-DA modelling practice strategies (see Fig. 3). It is important to stress here that most of the articles seldom inform the reader clearly on the decisions of all the aforementioned aspects. For this reason, one can see a number of dash bars (—) presented in ESI Table 1,† especially under the DR and model validation aspects.

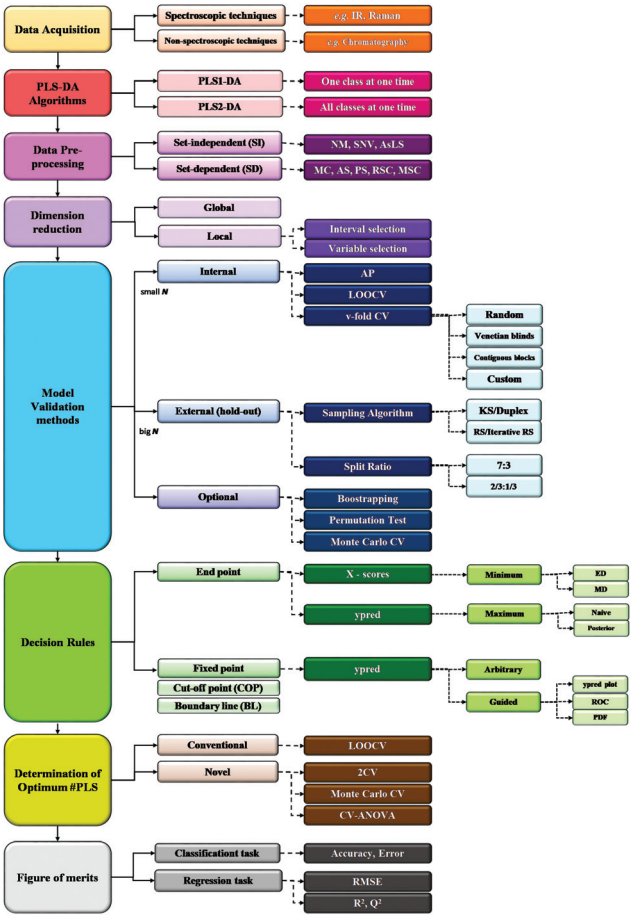

Fig. 3 Eight important practical aspects in PLS-DA modelling.[22–89] IR: infrared; NS: normalization; SNV: standard normal variate; AsLS: asymmetric least squares; MC: mean centring; AS: autoscaling; PS: Pareto scaling; RSC: robust scaling; MSC: multiplicative scatter correction; AP: autoprediction; LOOCV: leave-one-out cross validation; KS: Kennard–Stone sampling; RS: random sampling; ED: Euclidean distance; MD: Mahalanobis distance; ROC: receiver operating characteristic; PDF: probability density function; 2CV: repeated CV; RMSE: root mean squared error; R^2: coefficient of determination; Q^2: coefficient of prediction; N: data size; ypred: predicted values.

Next, each of the aspects will be first described in a general context, and then the related contemporary practice strategy in the context of PLS-DA is described by referring to ESI Table 1.† Eventually, rationales or knowledge gaps related to implementing the practice in colossal ($N > 1000$) and multi-class datasets are discussed. In fact, all the aspects listed in Fig. 3 have been discussed extensively in the literature, except the PLS-DA algorithm and the DR. For this reason, only the two said topics will be discussed in detail and others will only be briefly

explained. It is worth noting that the sequence of steps in Fig. 3 is not absolute but purely for illustrative purpose. In addition, the first practical aspect, *i.e.* data acquisition, is included in Fig. 3 for completeness. The primary concern in the following discussion is mainly central around PLS-DA modelling.

3.2 Data acquisition

Firstly, prior to statistical modelling, samples of concern need to be analysed with a particular instrumental technique to prepare a dataset. A wide array of cutting edge spectroscopic technologies has been proposed with each characterized by unique merits and pitfalls.[90] Recently, attenuated total reflectance-Fourier transform infrared (ATR-FTIR) spectroscopy is of special interest because it is non-invasive, non-destructive and fast in analysis time.[91]

3.3 Contemporary practices

Based on the brief survey (see ESI Table 1†), all the studies share one similar aim, *i.e.* to construct a PLS-DA prediction model for solving problem in a particular applied area, *e.g.* forensic science,[52,54] disease diagnosis,[24–27] and food authentication.[22,23,28–33] Overall, it is noted that PLS-DA was mostly used in modelling spectral-like data such as Raman and infrared spectra. Occasionally, data obtained using non-spectroscopic techniques like imaging[35,54] and chromatography[28,53] have also been used successfully in PLS-DA modelling.

3.4 Knowledge gaps

3.4.1 What makes PLS-DA so popular in spectral data?. The popularity of PLS-DA in modelling spectral data is sensible. First, spectral data are often of high dimensionality ($N < J$) and the variables (*i.e.* wavenumbers) tend to be correlated with each other (*i.e.* collinearity). Consequently, a classical discriminant method like LDA is not a feasible solution for spectral data.[92] Secondly, though PCA could be used to remediate the pitfalls of LDA associated with HD data, PLS-DA has often demonstrated better performances than PCA-LDA.[14,93,94] Recently, Prof. Brereton has highlighted the historical and technical factors that have partly contributed to the overwhelming popularity of PLS-DA in recent literature.[95]

3.5 PLS-DA algorithm

As discussed in section 2, the PLS-DA classification model can be constructed using either PLS1-DA or PLS2-DA algorithms, *i.e.* PLS1-DA models one class at a time; and PLS2-DA models more than one class simultaneously. Traditionally, for a binary classification problem ($G = 2$), PLS1-DA is usually the preferred choice of algorithm. On the other hand, a multi-class problem ($G > 2$) is often modelled using the PLS2-DA algorithm. If one wishes to employ the PLS1-DA algorithm in solving a multi-class problem, G PLS1-DA (that adopts one-*versus*-all framework) models are needed to accomplish the task.

3.6 Contemporary practices

Across the list of studies reviewed herein (see ESI Table 1†), only around 10% of them inform the readers about the type of algorithm (*i.e.* PLS1-DA or PLS2-DA) used in the study.[22,62,71,73,74,78,81,89] The other 90% of the studies hardly mentioned clearly the type of algorithm that was chosen in their studies, but simply stated 'PLS-DA' in the writings. In fact, more than 50% of the articles involve a multi-class problem.[22,26,28,30,39] Eventually, we think it is sensible to assume that those studies have employed the PLS2-DA algorithm since most statistical software would automatically select PLS2-DA when the number of classes is more than two. In other words, the community of users tend to choose the PLS2-DA algorithm to model the multi-class problem. In addition, it is also noted that several papers have just used the acronym 'PLS-DA' instead of 'partial least squares discriminant-analysis' in the title of the research papers.[54,63,78,80,89]

3.7 Knowledge gaps

3.7.1 Is PLS-DA the sole acronym used in the literature to refer to PLS algorithm applied in classification problem?. It is generally agreed that 'PLS-DA' is the most common acronym to refer to the PLS algorithm applied in classification problem. On top of this, other similar acronyms, *e.g.* DPLS, PLSDA, and PLS-LDA, have also appeared in the literature. So, do these acronyms refer to the same approach? In principle, the acronyms are referring to different classification methods that employ the PLS algorithm to produce new input variables from the original dataset, either the X-scores or the ypred (see Fig. 2).

Studies reported before the 2010s were found to show more interest on the X-scores rather than the ypred. For instance, Kemsley[96] compared PCA and PLS in reducing the dimension of an ATR-FTIR spectral dataset. LDA was then used to construct prediction models using the resulting PCA's scores and PLS's X-scores, respectively.[96] The proposed approach was called discriminant PLS in an article published one year later by the same author.[97] On the other hand, Marigheto *et al.*[98] adopted the same approach in discriminating extra-virgin and adulterated edible oil based on Raman and ATR-FTIR spectra and they termed the method PLS/LDA instead of discriminant PLS. The same approach was later shortened to PLS-LDA when rephrased by Tang *et al.*[99] On the other hand, Nocairi *et al.* also employed X-scores of PLS coupled with LDA in modelling near IR spectra and called the method PLS-DA.[93] Later, the PLS's X-scores were also being employed as input variables in other classification methods like logistic discriminant analysis[94] and quadratic discriminant analysis (QDA).[100] Both the approaches were not abbreviated in the original articles but were shortened to PLSda and PLSqda when rephrased by Mehmood *et al.*[5]

On the other hand, Ciosek *et al.* used PLS-DA to refer to a classification method that employed PLS's ypred (and not X-scores) in prediction and did not involve other discriminant methods (*e.g.* LDA or QDA) in translating the ypred into mean-

ingful class membership.[101] Following that, papers published after the 2010s tended to use PLS-DA to indicate a similar approach, *i.e.* translating class membership using the raw naïve ypred.[2,4] Occasionally, some papers have used PLSDA leaving the hyphen behind to refer to the same approach.[102,103] Sometimes, discriminant partial least squares/discriminant PLS (DPLS or D-PLS) are also employed to denote the method.[104–106]

Based on the discussion, it is clearly shown that PLS-DA has been used indiscriminately for the classification method that employed either X-scores[99] or ypred[101] as input variables. However, it is not our intention to advocate a standard acronym for the two different approaches. Technically, the acronym is a sort of personal preference. For this reason, we strongly advise researchers to read the paper thoroughly so as to understand the type of approach being employed by the authors prior interpretation in order to avoid drawing false conclusions.

3.7.2 Is PLS1-DA a better algorithm than PLS2-DA in modelling multi-class dataset?. To the best of our knowledge, there are only two empirical studies that have compared classification performance between PLS1-DA and PLS2-DA.[107,108] On one hand, Galtier *et al.*[107] demonstrated a significant difference between PLS1-DA and PLS2-DA in modelling two imbalanced 4- and 5-class multi-classification tasks (N = 36; 225). On the other hand, Serrano-Lourido *et al.*[108] has employed a balanced 3-group mass spectral dataset (N = 90) and reported an insignificant difference between the two algorithms. Both studies derived the conclusion in terms of model accuracy and have employed threshold-based DR to translate the class membership of unknown samples. However, Galtier *et al.*[107] determined the threshold arbitrarily whereas Serrano-Lourido *et al.*[108] employed a Receiver Operating Characteristic (ROC) curve to determine the most optimum cut-off values. Eventually, the 'significant difference' reported by Galtier *et al.*[107] could be unreliable since it is generally accepted that an arbitrary cut-off value is less likely to work well with an imbalanced dataset.[82] In addition, both the studies have not compared the algorithms using other aspects of model performances like model stability, robustness and parsimony. In other words, empirical differences between PLS1-DA and PLS2-DA are still unclear, especially in a complex dataset.

3.8 Data pre-processing (DP)

Due to the inherent limitations of most analytical instruments, the resulting data are seldom ready to be analysed to achieve the goal of analysis but would need to be pre-processed beforehand. However, to select the right DP method for the data at hand is not easy and could be a time consuming task.[109] Traditionally, DP methods can be categorized into set-dependent (*i.e.* 2-way) and set-independent (*i.e.* 1-way) methods. On one hand, the former methods apply column-wise operation that the whole dataset is employed to estimate the parameters. On the other hand, the latter methods are preferred over the former since they process the spectrum one by one and the required parameters are estimated from the individual spectrum itself (*i.e.* row-wise operation). Due to this, the third decision is about the DP strategy. Which DP method to use? How many DP methods shall be assessed? How to select the right DP method?

3.9 Contemporary practices

Based on the survey results as illustrated in ESI Table 1,† mean centring (MC), autoscaling (AS) and derivative using the Savitzky–Golay algorithm are the top three most used DP methods. It is also noted that most studies employed just one DP method to pre-process the data[22,26,28,30,32,38,41,43] whilst only a small fraction of users employed either a series of DP methods[24,27] or never pre-processed the data beforehand.[23,25] This indirectly shows the low awareness among applied scientists about the potential impact DP could have on the subsequent (PLS-DA) modelling output. In addition, the researchers seldom justified why a particular DP method was chosen.

3.10 Knowledge gaps

3.10.1 Is MC or AS really improving the PLS-DA model performance?. In principle, mean centring (MC) or autoscaling (AS) is performed to make sure all variables have comparable ranges and distributions.[110] In addition, for spectral-like data, especially IR spectra, such conditions are often fulfilled without AS or MC. As demonstrated in our preliminary study, a raw IR spectral dataset and the respective MC ones could perform equally well if the dataset is of high quality, as evaluated using the PCA-LDA method.[111] The rationales of applying MC prior to PLS habitually have been discussed elsewhere.[1] Theoretically, MC would not affect the model performance much if the dataset is not having any baseline problem. In the same work, we have also demonstrated that AS could induce varying impacts depending on the quality of the spectral dataset. However, we have not assessed the impacts of MC and AS in spectral dataset using PLS-DA and the dataset used in ref. 111 is rather small and simple.

3.10.2 What are the impacts of not pre-processing the dataset prior to PLS-DA modelling?. As shown by the survey, some researchers never consider any DP prior to PLS-DA modelling. Such practice could be partly explained by three rationales. On one hand, articles that published on the impacts of various DP methods in spectral dataset often exemplified using regression task.[112,113] On the other hand, the importance of DP prior to modelling is seldom highlighted in the literature. Moreover, to select the best DP method for the dataset on hand from the plethora of available DP methods is definitely a time consuming task. To the best of our knowledge, there has been no work attempt to address the potential impacts of not pre-processing the dataset prior to PLS-DA modelling.

3.11 Dimension reduction

Normally, the step after the DP is dimension reduction, which can be achieved *via* linear combination or variable selection. The potential merits of dimension reduction include: (a) reduced computational cost and time; (b) reduced risk of over-

fitting (*i.e.* improved model generalizability); and (c) better model interpretability. Assuming an IR spectral dataset, common linear combination techniques, *e.g.* PCA and PLS, would simply reduce the dimension of the spectral data into a smaller number of new axes; and the variable selection techniques might pick a number of discrete variables from the original data, *e.g.* to select a particular spectral region (*i.e.* interval selection), or to choose a number of discrete wavenumbers from the global IR spectral region. Contrary to linear combination, variable selection techniques come in many varieties.[114,115]

3.12 Contemporary practices

Based on the survey (see ESI Table 1†), most studies have employed the whole input region to produce a global model.[22–28,30–42] This practice is in line with a general belief that PLS-DA is capable of removing the analyte-irrelevant information by assigning relative low coefficient (*i.e.* loading) to the uninformative region. Occasionally, the global region is truncated based on prior knowledge of the researchers on the sample, *i.e.* interval selection.[29,43,44,52,69,82–84,88] A model that is built using a particular portion of the whole input region is normally known as a local model. Eventually, it can be seen here that PLS-DA modelling seldom involves variable selection approaches except for the interval selection approach.

3.13 Knowledge gaps

3.13.1 What are the merits of variable selection in PLS-DA modelling?. By re-searching the literature, we found that (at least) two papers have attempted to establish merits of genetic algorithms (GA), *i.e.* one popular VS method, in modelling the HD dataset using PLS-DA. Both studies concluded that GA-PLSDA models produced a lower error rate and better model robustness than the counterpart models that have not coupled with GA, *i.e.* PLS-DA models.[116,117] In practice, GA resembles another time consuming task that one needs to tune several parameters before reaching the most optimum outcomes.[118,119] On the other hand, Aliakbarzadeh *et al.*[120] investigated the abilities of five different VS methods (*i.e.* recursive partial least squares (rPLS), variable importance in projection (VIP), selectivity ratio (SR), significance multivariate correlation (sMC), and PLS loading weights) on the PLS-DA model, and concluded that all the models were comparable to each other in terms of prediction ability. However, the dataset was quite small (N = 83) and the authors have not compared the performances of the VS methods with respect to model stability or robustness except model accuracy.[120] In addition, to the best of our knowledge, reviews and manuscripts published on the impacts of variable selection techniques often focus only on PLS regression[121–123] rather than PLS-DA classification. In other words, the literature needs reports that: (a) evaluate the roles of variable selection in PLS-DA modelling according to varying model performances, *i.e.* model stability, parsimony and robustness; and (b) verify if the impacts of variable selection techniques being reported in the context of PLS regression[121–123] are also applicable in the context of PLS-DA classification task.

3.14 Model validation strategies

In order to construct a reliable prediction model, the choice of the model validation method and strategy is also important. A wide array of model validation methods has been proposed in the literature and none of them perfect.[124–136] In other words, it is not easy to estimate the true model performance because none of the model validation methods would appear optimum to all modelling settings (*e.g.* goal of analysis and nature of dataset).

3.15 Contemporary practices

Now, let's examine the favoured model validation methods being employed in estimating PLS-DA model performance (see ESI Table 1†). For the sake of brevity, the model validation methods are divided into three different groups: (a) internal methods, *e.g.* autoprediction (AP) and cross-validation (CV); (b) external testing (ET); and (c) optional methods such as bootstrapping and permutation test. Overall, it can be seen that internal methods are preferred over the external method, and the optional methods are the least used approach (see ESI Table 1†).

For internal validation, the most used variant is leave-one-out CV (LOOCV),[28–33,41–43] and then followed by 10-fold CV,[38,39,47,66,72,79,83,88] and 7-fold CV.[25,26,46,51,70] Based on ESI Table 1,† among those that have employed v-fold CV for model assessment, only several users have reported clearly the sampling framework used in running the v-fold CV, *i.e.* random sampling,[22] stratified sampling,[50] contiguous blocks CV,[54,75] and Venetian blinds CV.[48,49,51,52,66,78,88] In practice, it seems relevant to just assume that the rest have employed the random splitting mode since it is the default setting in most statistical software. In addition, it is noteworthy to mention that several users that employed random v-fold CV have repeated the procedures several times.[40,47,53,72,79] In one case, the authors employed custom CV that the test samples were pre-selected according to specific criteria.[50] Other than that, some studies employed CV that integrated with the iterative resampling method, such as repeated CV,[53] Monte Carlo CV (MCCV),[58] and CV-ANOVA.[60]

On the other hand, the favoured sampling algorithm and split ratio being employed in running the ET are also recorded. In overall, Kennard–Stone[35,48,57,59] and random sampling[22,23,28,47,54,55] are the two most used sampling algorithms and the favoured split ratios are 7 : 3 and 2/3 : 1/3. In addition, only a few studies have employed repeated random sampling.[34,41,58,65,82,85] Last but not least, very few users have employed computational intensive methods, *i.e.* bootstrapping[79,89] and permutation test[24,25,27,36] to further evaluate the performance of the PLS-DA model. Nonetheless, it is clearly demonstrated in ESI Table 1† that the permutation test was applied relatively more frequently than bootstrapping.

3.16 Knowledge gaps

3.16.1 Which model validation method shall be used? Internal validation or external validation?. In practice, the choice relies heavily on the size of the data that a small dataset is best validated using an internal validation method like CV;[124,125] and a sufficiently big dataset shall be validated using an external testing approach.[126] In other words, the PLS-DA model being constructed using the colossal dataset ($N > 1000$) is expected to be the best validated using ET.

3.16.2 Is 7-fold CV a more feasible choice over the 5- or 10-fold CV?. It was a bit surprising that 7-fold CV has appeared to be the new favourable variant of *v*-fold CV on top of the classical choices, *i.e.* 5- or 10-fold CV.[128,129] Theoretically, variance, bias and computational burden are the three major issues one shall consider when deciding on the right number of folds to run *v*-fold CV. As has been demonstrated by Hastie *et al.*,[130] the sample sizes could significantly affect the bias of the resulting estimate of prediction error. On the other hand, as the number of *v* is reduced the associated bias is increased but both the variance and computational burden are reduced. Eventually, 7-fold CV resembles a compromise option between the 5- and 10-fold CV in all the three terms.

3.16.3 How shall one split the dataset in *v*-fold CV? Random or systematic sampling framework?. For *v*-fold CV, different settings have been proposed to split the dataset into different folds. Basically, random splitting is easy to implement but can be subjected to unforeseen bias that could result in falsely optimistic model performance. In order to reduce the unforeseen bias, we think the most feasible way is to repeat the random sampling multiple times.[40,47,53,72,79] Otherwise, CV that employed systematic sampling, *i.e.* contiguous blocks and Venetian blinds, can be another viable choice of method. Both the methods are especially valuable for imbalanced, colossal and multi-class data; and the selection depends on how the classes are arranged in the dataset. More details of the contiguous blocks and Venetian blinds are available elsewhere.[4,137] Otherwise, if the researcher knows the samples very well, custom splitting could be another viable option. When the PLS-DA prediction model is formed using a complex dataset, it seems to be much manageable to conduct *v*-fold CV that employs systematic resampling (*i.e.* contiguous blocks and Venetian blinds) in order to ensure the reliability of the resulting model estimate.

3.16.4 What are the advantages and disadvantages of Kennard–Stone against random sampling?. In contrast to Kennard–Stone sampling that ensures uniform distribution of samples between training and test sets,[138] random sampling does not employ any principle in splitting the dataset. For this reason, a good practice is to repeat the random sampling multiple times (*i.e.* iterative random sampling) in order to reduce unforeseen bias caused by the non-uniform sampling.[34,41] However, we noticed that only several studies have performed this good practice.[34,41,58,65,82,85] The pros and cons of the Kennard–Stone and the random sampling in the context of regression analysis have been presented elsewhere.[131] Theoretically, iterative random sampling is more likely than the Kennard–Stone algorithm to produce test samples that resemble closely the unforeseen future events. However, in practice, when involving a colossal dataset ($N > 1000$), it is unclear if it is worth investing more time on estimating model performance using the iterative random sampling against the Kennard–Stone algorithm.

3.16.5 Is it important to validate a PLS-DA model using permutation test and bootstrapping?. In practice, it has been demonstrated that PLS-DA could produce a good model even with null data.[20,96] The probability of an unrealistic separation (no meaningful relationship) tends to increase with the number of variables.[2] For this reason, a good practice is to determine the presence of true relationship between X and Y *via* the permutation test.[136] On the other hand, bootstrapping is usually employed for determining the robustness of the model.[135] However, both methods are time and computationally demanding as well as seldom readily available in the statistical software. The difficulty of running both tests on the colossal dataset ($N > 1000$) is expected to increase. However, due to the supervised nature of PLS-DA,[2–4] we do believe that the permutation test is an essential step in PLS-DA modelling such that the calculation time would be subordinate.

3.16.6 What are the differences between bootstrapping and iterative based cross validation methods?. Besides the bootstrapping technique, double cross validation (2CV),[20] repeated 2CV,[139] and Monte Carlo (leave-many-out) random subsampling validation[140] have also been demonstrated to be a viable choice of techniques in confirming predictive power of the PLS-DA model. In contrast to the permutation test that aims at assessing the significance of the PLS-DA model,[20,27] bootstrapping, 2CV, repeated 2CV and MCCV are quite similar to each other and all involve an iterative random resampling framework to derive the uncertainty associated with the model performances.[20,79,135,139,140] The differences between bootstrapping and Monte Carlo based simulation (including MCCV) have been reported by several researchers.[126,141–143]

3.16.7 What are the impacts of the type of DP method on the model validation strategy?. Basically, the choice of DP method could affect the reliability of the model validation strategy. The problem arose when one decided to use a 2-way DP method to pre-process the dataset. In practice, the 2-way DP method is often performed on an entire experimental dataset. This can have a serious consequence when one assesses the quality of the model using an external testing approach. By right, if the PLS-DA model is decided to be pre-processed with a 2-way DP method and validated externally, then the parameter (of DP) shall be estimated from the training set alone instead of the whole data and the test set is then pre-processed using the same parameter(s). However, such safeguard is seldom implemented by most researchers. The issue is of particular concern since it has been demonstrated in the earlier section that two of the three most favourable DP methods (*i.e.* MC and AS) are indeed 2-way DP methods. In practice, the same caution shall also be applied to the *v*-fold CV method that the test set in every cycle should be pre-pro-

cessed independent from the respective training set.[141,144] The issue becomes complicated if an iterative-based model validation method, *e.g.* bootstrapping or permutation test, is employed for validating the model. Otherwise, the estimated model performance could be falsely optimistic.

3.17 Decision rules (DRs)

Fifth, if the model is intended for prediction purpose, one needs to decide on the type of decision rule to be employed in PLS-DA prediction. However, this aspect is often overlooked by the community of users. Given a set of training data that belongs to G classes, each new sample is reduced into G predicted values ($\hat{y}$), estimated *via* the G class-specific regression coefficient (see Fig. 2). Ideally, the predicted value is either 0 or +1, corresponding to perfect membership prediction. But, since PLS is inherently designed for continuous output variables (*i.e.* regression task), $\hat{y}$ can never be an integer but takes any values between '0' to '1'. For this reason, a DR is required to translate $\hat{y}$ into meaningful class membership accurately.

In practice, the variety of DRs can be divided into two major groups: (a) end point and (b) fixed point. In the former group, the X-scores or ypred could be used together with a particular approach to determine the class membership of the test sample. On one hand, X-scores could be converted into a particular distance metric, *e.g.* Euclidean distance (ED) or Mahalanobis distance (MD), where the sample would be assigned to the class that exhibited the minimum distance. On the other hand, the ypred could be used in its raw (naïve) form or manipulated into a form of posterior probability *via* a particular probability density function. The fixed point based approaches employ only the ypred and could involve either one fixed point (*i.e.* cut-off point) or two fixed points (*i.e.* boundary line). The optimum point(s) can be determined either arbitrarily or according to particular diagnostic tools, *e.g.* ypred plot, receiver operating characteristic curve (ROC) and probability density function. However, most of the time, the cut-off point is determined to be halfway between the means of the two groups.

3.18 Contemporary practices

Based on the survey (see ESI Table 1†), it is noted that the three most popular DRs are: (a) naive (maximum value, Max), (b) cut-off point, and (c) boundary line. Basically, many users rarely stated the details of DR in the writings. Consequently, it seems logical to assume that they have actually employed the naïve DR (Max), since it is the simplest rule and is the default parameter in most statistical software. Several studies have employed cut-off point based DR with the help of varying diagnostic tools, *e.g.* Y-score plot,[44,52,54] ROC[36] and probability density function,[53] in estimating the optimum points. The least used DR is the boundary line based DR which would assign an unknown sample to none of the groups, *i.e.* unassigned. The low popularity of the boundary line could be because of the relatively complex computation step. Consequently, the survey seems to suggest that the DR has not received considerable attention which is partly caused by the lack of understanding of the pitfalls of the naïve DR.

3.19 Knowledge gaps

3.19.1 What are the pitfalls of the naïve DR?. To the best of our knowledge, there are only three papers that have compared empirical performances of naïve DR (Max) against other novel DRs.[101,145,146] Basically, two key ideas can be drawn from the studies. First, ypred tends to outperform the respective X-scores. Secondly, a novel DR including extra terms in the ypred tends to outperform the naïve ypred. The excellent performance of the novel DR is expected since a more flexible method tends to produce a less bias model.[147]

Next, theoretical differences between the naïve and selected novel DRs will be illustrated on a hypothetical dataset, as depicted in Fig. 4. Assuming that we have constructed a PLS-DA model using a binary and balanced hypothetical dataset that consists of 20 samples with each group composed of 10 samples, both the groups are respectively recoded into dummy code, *i.e.* +1 and 0. Specifically, let y_1, y_2, ..., y_{20} be a set of predicted values derived *via* internal validation from the same dataset. Next, let us see how the different DRs would perform under three different scenarios. The best demarcation line between the two groups is indicated by red dash lines in Fig. 4.

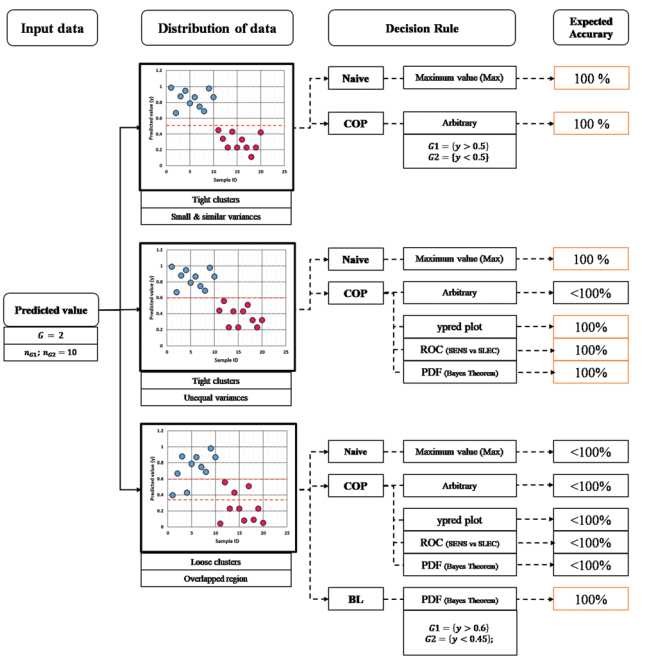

Fig. 4 Expected accuracy of PLS-DA models as predicted *via* different decision rules. The models are constructed using three different balanced binary hypothetical data that differ in terms of sample distributions and variances. G: number of groups; COP: cut-off point; G1: group 1; G2: group 2; y: predicted value; SENS: sensitivity; SLEC: selectivity; PDF: probability density function; ROC: receiver operating characteristic; BL: boundary line.

First, the resulting predicted values presented two tight clusters (*i.e.* compact groups) and each exhibited small and similar variance (spread of data). Both groups are well separated and the best demarcation line is located at 0.5 (see red dash line in top figure in Fig. 4). This scenario is perfect and one would achieve 100% accuracy using the naive DR. Alternatively, one can determine an arbitrary threshold (*i.e.* cut-off point), *i.e.* $\frac{1+0}{2} = 0.5$, and still can expect a 100% prediction rate.

Next, let's further complicate the scenario by making both the groups show unequal variances, and the best demarcation line is now shifted to 0.6 (red dash line in the middle figure in Fig. 4). In this example, the naïve DR would still be able to achieve 100% accuracy whilst arbitrary threshold (*i.e.* cut-off point) cannot predict all the samples correctly since the optimum demarcation line has shifted to 0.6 instead of 0.5. Alternatively, the optimum threshold could be determined with the help of visualization tools: (a) ROC,[108] (b) Y-score plot,[44,52] and (c) probability density function/plot (see Fig. 5A).[53,146]

Now let's propose an even worst scenario that both the groups presented loose clusters (*i.e.* dispersed groups) and shared a confusion region in which both the classes are overlapped (see Fig. 4). In reality, an imbalanced or inhomogeneous real-world datasets would tend to present such scenario. Apparently, the naïve and also cut-off point based DRs are no more able to achieve 100% accuracy. In this case, one can opt to determine the range of predicted values specific to each class, so any sample falling outside the interval will be labelled as an unclassified sample or otherwise. In our example, there are five unassigned samples as estimated using the boundary line based DR. Just like the cut-off point, one can employ different diagnostic plots in finding the most optimum range of values for each class (see Fig. 5B).

Basically, the discussion presented above relied only on the first two PLS components. In practice, it is important to note that all the DRs could possibly achieve a zero error of prediction if supported by a sufficient number of PLS components. The question now is which DR could produce perfect performance with the minimum number of PLS components.

3.19.2 What is the cost of using more complicated DR over the naïve DR?. In practice, the naïve DR could be optimized *via* different diagnostic tools, *i.e.* ypred plot, ROC and probability density function. In principle, the former two tools are simpler to conduct than the latter one. Basically, the ypred plot is the most naïve plot (as shown in Fig. 4) and might not work well when the number of classes increases. On the other hand, ROC[148] reproduces similar information in a ypred plot from another perspective and possibly eases the load to determine the optimum threshold for multi-class problem. However, both the approaches do not provide allowance to reduce bias that could occur from unequal group sizes in the dataset.

In contrast to ROC or ypred plot, the probability density function based DR provides the highest modelling flexibility. Two favoured approaches for estimating the probability density function are the parametric Gaussian method,[104,105] and the non-parametric Kernel method.[53] The former method requires the input data (*i.e.* ypred) to follow the Gaussian distribution whilst the latter method does not presume any form of distribution. Once the probability density function has been established, the bias occurred from unequal group sizes could be resolved by using customized prior probability incorporated into the probability density function *via* the Bayes theorem.[44,149] Nonetheless, to choose the right priors for the data at hand is not easy and the customized prior probabilities shall take into account the underlying populations which are often not known by the researchers.[150] Alternatively, one can also choose not to include any priors, *i.e.* equal prior probability. However, the PDF based DR is computationally intensive and a small dataset needs to employ resampling techniques such as MCCV[146] to produce reliable (ypred) distribution. Otherwise, one can just use the raw ypred if the dataset is sufficiently large. Eventually, the reliability of the resulting density plot heavily depends on the inherent properties of data, *i.e.* representativeness, and group and sample sizes.

In essence, we can see that naïve and cut-off points are simpler than the probability density function based DR but might not work well with the contemporary complex dataset, *i.e.* imbalanced and multi-class dataset. On the other hand, the probability density function based DR is expected to work well in the complex dataset but at the cost of increased risk of overfitting.[147] To the best of our knowledge, the performances of the naïve DR and probability density function based DR have not been compared in terms other than model accuracy. Eventually, we think that it is worth presenting an empirical study that compares the naïve and probability density function based DRs using a complex dataset by referring to varying aspects of a prediction model, *e.g.* model stability and model parsimony.

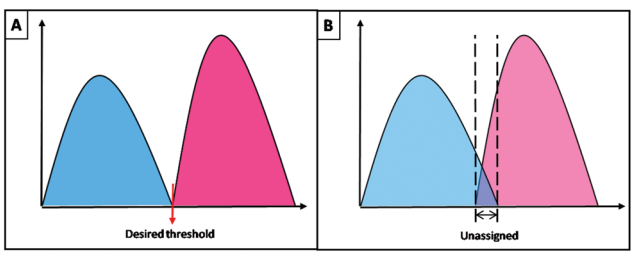

Fig. 5 Two possible scenarios could be presented when the predicted values are plotted using non-Gaussian peaks: (A) non-overlapped distributions. (B) Overlapped distributions.

3.20 Tuning of model complexity

In PLS-DA modelling, one needs to determine the number of PLS components to be retained in the formation of the desired prediction model.

3.21 Contemporary practices

Our survey results (see ESI Table 1†) show that only a very small fraction of the users did mention clearly how the optimum number of PLS is determined. Basically, LOOCV is the most preferred approach.[28,41,43,45,63] In contrast to these studies that discussed the model performances using only the optimum number of PLS components, it is worth mentioning that several researchers have presented the model performances along a pre-set range of PLS components.[23,38,67,87] Last but not least, only one work being reviewed in this article has employed the time consuming approach, *i.e.* MCCV, to determine the parameters of the PLS-DA method.[58]

3.22 Knowledge gaps

3.22 1 Is LOOCV a reliable approach for tuning of model complexity? What are the more viable alternatives?. Despite LOOCV having far less bias, it could be infeasible as sample size increases.[128] Sometimes, error being estimated using the LOOCV could be unreliable.[126,141,143,151–154] Consequently, alternative approaches like MCCV,[154,155] 2CV,[20,151] and repeated double CV (rdCV)[139,143] have been proposed in assessing the model complexity thoroughly. The merits of rdCV have been determined by assessing its model performance over the counterpart models being constructed using original variables, and using a subset of the variables selected by GA and MCCV.[143] Krakowska *et al.*[156] have reported the Monte Carlo validation framework for the PLS-DA model extended with VS methods in deriving the most optimum model complexity. On the other hand, differences between LOOCV and MCCV in the context of PLS regression have been demonstrated by Xu and Liang.[154] Theoretically, techniques that involve a random resampling method, *e.g.* 2CV, MCCV and rdCV, are expected to outperform LOOCV in determining the right model complexity.

3.23 Figures of merit

Figures of merit, known also as metric or model estimates, are values being estimated from a prediction model by using a particular model validation method. In practice, varying figures of merit have been proposed to explain the performance of a regression or classification model. Basically, the choice of figures of merit depends on the goal of analysis and inherent nature of the data set.

3.24 Contemporary practices

Two most used figures of merit in PLS-DA modelling are classification error rate,[22,23] and root mean squared error (RMSE).[29–33] Other figures of merit also in use are sensitivity and specificity,[29,30,44] coefficient of determination of model fitting (R^2), and prediction (Q^2),[24–28] area under curve (AUC) and ROC.[36,39,40]

3.25 Knowledge gaps

3.25.1 Is RMSE suitable to describe the performance of a PLS-DA model?. Researchers habitually report the performance of a classification model in terms of error rate or classification accuracy. In contrast to other discriminant methods like Fisher's Linear Discriminant Analysis (LDA),[157] and SIMCA,[158] PLS-DA is derived from a regression algorithm, *i.e.* PLS. Due to this reason, figures of merit that are common in the regression model, *e.g.* RMSE, R^2, and Q^2 have been employed to describe a PLS-DA model (see ESI Table 1†). A primary difference between RMSE and error rate is that the former uses the raw predicted value to estimate the classification ability of the model, whilst the latter employs the translated version of the predicted value (*i.e.* a particular DR is used to convert $\hat{y}$ into meaningful class membership). Mathematically, RMSE will not be affected by the DR. However, in practice, RMSE does not seem to be relevant in the context of classification.[102] On the other hand, Gromski *et al.* have highlighted that R^2 and Q^2 are invalid to explain a categorical model.[19]

3.26 General remarks

In the contemporary big data era where data sizes grow rapidly attributed to the rapid development in instrumental technology, the real-world data are evolving into a more complex form. However, it is noted here that PLS-DA has not been explored in modelling complicated datasets. Many studies have employed rather simple datasets with a moderate number of samples ($N < 200$) on a small number of groups ($G < 5$). Despite most studies often emphasize that their results are readily extended to other similar studies, but with a bigger dataset (and involved more number of groups), the problem becomes more complicated and some technical aspects that are proved to be valid in binary data might not fit to multi-class data. Eventually, the empirical statistical literature lacks exemplar work that demonstrates the potential of PLS-DA in modelling colossal, imbalanced and multi-class data.

In addition, most of the studies tend to report only model accuracy and seldom give attention to other model performances, *e.g.* model stability,[128,159] parsimony,[160] and fitting.[161] Since most of the works or the ultimate aim of a particular statistical algorithm is indeed to provide a solution in real-world applications, the researchers shall also have evaluated the models in terms of model stability or parsimony which are of paramount importance in deriving the potential of the model in future use.

4 Conclusion

PLS-DA is a powerful algorithm for predictive and also descriptive modelling. However, the price to pay for the diverse possibilities of outcomes is the optimization of a wealth of parameters. In this paper, we have first briefly explained the theoretical background of the PLS-DA algorithm. Following that, the PLS-DA predictive modelling practice strategies are presented and critically discussed. Basically, the primary concerns of this work are devoted to the PLS-DA algorithm and the decision rules since both are seldom being discussed in depth in the contemporary literature. In conclusion, more empirical studies

are required to refine the PLS-DA modelling practice strategies especially in complex datasets, *i.e.* high dimensional, multi-class, imbalanced and colossal, that resemble closely the future real-world problems.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

The research received funding from the CRIM UKM (Grant No. GUP-2017-043) and the Malaysian Ministry of Education (Grant No. FRGS/2/2013/ST06/UKM/02/1). We would also like to acknowledge the contributions of the anonymous reviewers for their constructive comments and ideas. Special thanks goes to Wan Nur Syazwani Wan Mohamad Fuad for her editorial support.

References

1 R. G. Brereton, *Analyst*, 2000, **125**, 2125–2154.
2 R. G. Brereton and G. R. Lloyd, *J. Chemom.*, 2014, **28**, 213–225.
3 M. L. Barker, *Partial least squares for discrimination, statistical theory and implementation*, LAP LAMBERT Academic Publishing, Germany, 2015.
4 D. Ballabio and V. Consonni, *Anal. Methods*, 2013, **5**, 3790–3798.
5 T. Mehmood and B. Ahmed, *J. Chemom.*, 2016, **30**, 4–17.
6 R. G. Brereton, *Chemometrics for pattern recognition*, John Wiley & Sons Ltd, Chichester, England, 2009.
7 N. Kumar, A. Bansal, G. S. Sarma and R. K. Rawal, *Talanta*, 2014, **123**, 136–199.
8 L. Wu, B. Du, Y. V. Heyden, L. Chen, L. Zhao, M. Wang and X. Xue, *TRAC, Trends Anal. Chem.*, 2017, **86**, 25–38.
9 J. Ahlinder, A. Nordgaard and S. W. Lindstrom, *J. Chemom.*, 2015, **29**, 267–276.
10 M. Sattlecker, N. Stone and C. Bessant, *TRAC, Trends Anal. Chem.*, 2014, **59**, 17–25.
11 J. Trevisan, P. P. Angelov, P. L. Carmichael, A. D. Scott and F. L. Martin, *Analyst*, 2012, **137**, 2302–2312.
12 C. Serrano-Cinca and B. Gutierrez-Nieto, *Decis. Support Syst.*, 2013, **54**, 1245–1255.
13 L. C. Soares, J. O. Alves, L. A. Linhares, F. B. E. Filho and M. P. F. Fontes, *Microchem. J.*, 2017, **133**, 258–264.
14 M. Barker and W. Rayens, *J. Chemom.*, 2003, **17**, 166–173.
15 L. C. Lee, C.-Y. Liong and A. A. Jemain, in *Seminar Kebangsaan Institut Statistik Malaysia ke-11 (SKISM-XI) 2017*, UKM, 2017.
16 L. C. Lee, C.-Y. Liong and A. A. Jemain, *AIP Conf. Proc.*, 2016, **1750**, 060016.
17 L. C. Lee, C.-Y. Liong and A. A. Jemain, in *2017 National Forensic Science Symposium (NFSS 2017)*, Forensic Science Society of Malaysian, 2017.
18 M. Grootveld, in *Metabolic Profiling, Disease and Xenobiotics*, Royal Society of Chemistry, England, 2012, pp. 1–34.
19 P. S. Gromski, H. Muhamadali, D. I. Ellis, Y. Xu, E. Correa, M. L. Turner and R. Goodacre, *Anal. Chim. Acta*, 2015, **879**, 10–23.
20 J. A. Westerhuis, H. C. J. Hoefsloot, S. Smit, D. J. Vis, A. K. Smilde, E. J. J. van Velzen, J. P. M. Duijnhoven and F. A. van Dorsten, *Metabolomics*, 2008, **4**, 81–89.
21 E. Szymanska, E. Saccenti, A. K. Smilde and J. A. Westerhuis, *Metabolomics*, 2012, **8**, S3–S16.
22 M. L. Amodio, F. Ceglie, M. M. A. Chaudhry, F. Piazzolla and G. Colelli, *Postharvest Biol. Technol.*, 2017, **125**, 112–121.
23 B. Yang, L. Yao and T. Pan, *Engineering*, 2017, **9**, 181–189.
24 J.-L. Wu, C.-X. Zhou, P.-J. Wu, J. Xu, Y.-Q. Guo, F. Xue, A. Getachew and S.-F. Xu, *PLoS One*, 2017, e0175573.
25 L. Vitova, Z. Tuma, J. Moravec, M. Kvapil, M. Matejovic and J. Mares, *BMC Nephrol.*, 2017, **18**, 112.
26 S. G. Snowden, A. A. Ebshiana, A. Hye, Y. An, O. Pletnikova, R. O'Brien, J. Troncoso, C. Legido-Quigley and M. Thambisetty, *PLoS Med.*, 2017, **14**, e1002266.
27 R. K. Sharma, K. Mishra, A. Farooqui, A. Behari, V. K. Kapoor and N. Sinha, *Inflammation Res.*, 2017, **66**, 97–105.
28 Q. Peng, X. Xu, C. Shen, R. Tian, B. Hu, X. Li, H. Zou, F. Chen, J. Wang, J. Jin, B. Li and G. Xie, *Innov. Food Sci. Emerg. Technol.*, 2017, **44**, 212–216.
29 M. K. Nieuwoudt, S. E. Holroyd, C. M. McGoverin, M. C. Simpson and D. E. Williams, *Appl. Spectrosc.*, 2017, **71**, 308–312.
30 A. R. Martins, M. Talhavini, M. L. Vieira, J. J. Zacca and J. W. B. Braga, *Food Chem.*, 2017, **229**, 142–151.
31 F. Mabood, J. Hussain, M. M. O. Al Nabhani, S. A. Gilani, S. Farooq, Z. Naureen, F. Jabeen, M. Ahmed, Z. Hussain and A. Al-Harrasi, *J. Adv. Dairy Res.*, 2017, **5**, 1000167.
32 F. Mabood, F. Jabeen, M. Ahmed, J. Hussain, S. A. A. Al Mashaykhi, Z. M. A. Al Rubaiey, S. Farooq, R. Boque, L. Ali, Z. Hussain, A. Al-Harrasi, A. L. Khan, Z. Naureen, M. Idrees and S. Manzoor, *Food Chem.*, 2017, **221**, 746–750.
33 X. Li, K. Xu, Y. Zhang, C. Sun and Y. He, *PLoS One*, 2017, **12**, 0169430.
34 A. Khoshmanesh, D. Christensen, D. Perez-Guaita, I. I. Ormaetxe, S. L. O'Neill, D. McNaughton and B. R. Wood, *Anal. Chem.*, 2017, **89**, 5285–5293.
35 K. D. T. M. Milanez, T. C. A. Nobrega, D. S. Nascimento, M. Insausti and M. J. C. Pontes, *Microchem. J.*, 2017, **133**, 669–675.
36 H. Jorgensen, A. S. Hill, M. T. Beste, M. P. Kumar, E. Chiswick, P. Fedorcsak, K. B. Isaacson, D. A. Lauffenburger, L. G. Griffith and E. Qvigstad, *Fertil. Steril.*, 2017, **107**, 1191–1199.
37 S. M. Azcarate, R. Gil, P. Smichowski, M. Savio and J. M. Camina, *Microchem. J.*, 2017, **130**, 1–6.

38 M. Garriga, S. Romero-Bravo, F. Estrade, A. Escobar, I. A. Matus, A. del Pozo, C. A. Astudillo and G. A. Lobos, *Front. Plant Sci.*, 2017, **8**, 280.
39 A. P. DeFilippis, P. J. Trainor, B. G. Hill, A. R. Amraotkar, S. N. Rai, G. A. Hirsch, E. C. Rouchka and A. Bhatnagar, *PLoS One*, 2017, **12**, e0175591.
40 P. D. Boccio, F. Perrotti, C. Rossi, I. Cicalini, S. Di Santo, M. Zucchelli, P. Sacchetta, D. Genovesi and D. Pieragostino, *Adv. Radiat. Oncol.*, 2017, **2**, 118–124.
41 M. Manfredi, E. Barberis and E. Marengo, *Appl. Phys. A*, 2017, **123**, 35.
42 K. Georgouli, J. M. D. Rincon and A. Koidis, *Food Chem.*, 2017, **217**, 735–742.
43 M. Kharbach, R. Kamal, M. Bousrabat, M. A. Mansouri, I. Barra, K. Alaoui, Y. Cherrah, Y. V. Heyden and A. Bouklouze, *Chemom. Intell. Lab. Syst.*, 2017, **162**, 182–190.
44 V. H. J. M. dos Santos, A. S. Ramos, J. P. Pires, P. de M. Engelmann, R. V. Lourega, J. M. M. Ketzer and L. F. Rodrigues, *Chemom. Intell. Lab. Syst.*, 2017, **161**, 70–78.
45 J. Peng, K. Song, H. Zhu, W. Kong, F. Liu, T. Shen and Y. He, *Sci. Rep.*, 2017, **7**, 44551.
46 L. Bogdanovska, A. P. Panovska, M. Popovska, A. Dimkitrovska and R. Petkovska, *Saudi Pharm. J.*, 2017, **25**, 1022–1031.
47 F. J. Cuevas, J. M. Moreno-Rojas and M. J. Ruiz-Moreno, *Food Chem.*, 2017, **221**, 1930–1938.
48 M. A. C. Reed, R. Singhal, C. Ludwig, J. B. Carrigan, D. G. Ward, P. Taniere, D. Alderson and U. L. Gunther, *Neoplasia*, 2017, **19**, 165–174.
49 R. Rios-Reina, S. Elcoroaristizabal, J. A. Ocana-Gonzalez, D. L. Garcia-Gonzalez, J. M. Amigo and R. M. Callejon, *Food Chem.*, 2017, **230**, 108–116.
50 M. V. L. Soares, E. G. A. Filho, L. M. Silva, E. H. Novotny, K. M. Canuto, N. J. Wurlitzer, N. Narain and E. S. de Brito, *Food Chem.*, 2017, **219**, 1–6.
51 M. Vermathen, M. Marzorati, G. Diserens, D. Baumgartner, C. Good, F. Gasser and P. Vermathen, *Food Chem.*, 2017, **233**, 391–400.
52 J. Manheim, K. C. Doty, G. McLaughlin and I. K. Lednev, *Appl. Spectrosc.*, 2016, **70**, 1109–1117.
53 M. Alewijn, H. V. D. Voet and S. V. Ruth, *J. Food Compos. Anal.*, 2016, **51**, 15–23.
54 L. Valderrama and P. Valderrama, *Chemom. Intell. Lab. Syst.*, 2016, **156**, 188–195.
55 F. B. de Santana, L. C. Gontijo, H. Mitsutake, S. J. Mazivila, L. M. de Souza and W. B. Neto, *Food Chem.*, 2016, **209**, 228–233.
56 D. Melucci, A. Bendini, F. Tesini, S. Barbieri, A. Zappi, S. Vichi, L. Conte and T. G. Toschi, *Food Chem.*, 2016, **204**, 263–273.
57 P. H. G. D. Diniz, M. F. Barbosa, K. D. T. de Melo Milanez, M. F. Pistonesi and M. C. U. de Araujo, *Food Chem.*, 2016, **192**, 374–379.
58 S. Hou, *J. Chemom.*, 2016, **30**, 663–681.
59 L. C. de Carvalho, C. L. M. Morais, K. M. G. de Lima, L. C. Cunha Jr., P. A. M. Nascimento, J. B. de Faria and G. H. de A. Teixeira, *Anal Methods*, 2016, **28**, 5658–5666.
60 M. Zotti, S. A. D. Pascali, L. D. Coco, D. Migoni, L. Carrozzo, G. Mancinelli and F. P. Fanizzi, *Food Chem.*, 2016, **196**, 601–609.
61 P. H. Rodrigues Jr., K. de Sa Oliveira, C. E. R. de Almeida, L. F. C. de Oliveira, R. Stephani, M. da S. Pinto, A. F. de Carvalho and I. T. Perrone, *Food Chem.*, 2016, **196**, 584–588.
62 A. Hirri, A. Balouki and A. Oussama, *Basic Res. J.*, 2016, **5**, 103–108.
63 W. Liu, Z. Sun, J. Chen and C. Jing, *J. Spectrosc.*, 2016, 1603609.
64 Y. Li, J. Zhang, T. Li, H. Liu and Y. Wang, *PLoS One*, 2016, **11**, e0168998.
65 E. Borras, J. Ferre, R. Boque, M. Mestres, L. Acena, A. Calvo and O. Busto, *Food Chem.*, 2016, **203**, 314–322.
66 S. Shrestha, M. Knapic, U. Zibrat, L. C. Deleuran and R. Gislum, *Sens. Actuators, B*, 2016, **237**, 1027–1034.
67 A. Racz, D. Bajus, M. Fodor and K. Heberger, *Chemom. Intell. Lab. Syst.*, 2016, **151**, 34–43.
68 L. Lenhardt, R. Bro, I. Zekovic, T. Dramicanin and M. D. Dramicanin, *Food Chem.*, 2015, **175**, 284–291.
69 D. M. I. Ho, A. E. Jones, J. Y. Goulermas, P. Turner, P. Turner, Z. Varga, L. Fongaro, T. Fanghanel and K. Mayer, *Forensic Sci. Int.*, 2015, **251**, 61–68.
70 Y. Wang, L. Xu, H. Shen, J. Wang, W. Liu, X. Zhu, R. Wang, X. Sun and L. Liu, *Sci. Rep.*, 2015, **5**, 18926.
71 S. J. Mazivila, H. Mitsutake, F. B. D. Santana, L. C. Gontijo, D. Q. Santos and W. B. Neto, *J. Braz. Chem. Soc.*, 2015, **26**, 642–648.
72 X. Shao, H. Li, N. Wang and Q. Zhang, *Sensor*, 2015, **15**, 26726–26742.
73 A. Hirri, A. Boulli and A. Oussama, *Int. J. Chem. Mater. Environ. Res.*, 2015, **2**, 30–36.
74 S. Moncayo, S. Manzoor, F. Navarro-Villoslada and J. O. Caceres, *Chemom. Intell. Lab. Syst.*, 2015, **146**, 354–364.
75 R. Calvini, A. Ulrici and J. M. Amigo, *Chemom. Intell. Lab. Syst.*, 2015, **146**, 503–511.
76 F. S. L. Borba, R. S. Honorato and A. de Juan, *Forensic Sci. Int.*, 2015, **249**, 73–82.
77 H. Chen, Z. Lin, H. Wu, L. Wang, T. Wu and C. Tan, *Spectrochim. Acta, Part A*, 2015, **135**, 185–191.
78 B. G. Botelho, N. Reis, L. S. Oliveira and M. M. Sena, *Food Chem.*, 2015, **181**, 31–37.
79 P. S. Gromski, E. Correa, A. A. Vaughan, D. C. Wedge, M. L. Turner and R. Goodacre, *Anal. Bioanal. Chem.*, 2014, **406**, 7581–7590.
80 V. A. G. da Silva, M. Talhavini, I. C. F. Peixoto, J. J. Zacca, A. O. Maldaner and J. W. B. Braga, *Microchem. J.*, 2014, **116**, 235–243.
81 J. M. D. Paulo, J. E. M. Barros and P. J. Barbiera, *Energy Fuels*, 2014, **28**, 4355–4361.

82 G. Tang, K. Tian, X. Song, Y. Xiong and S. Min, *Spectrochim. Acta, Part A*, 2014, **121**, 678–684.
83 O. Devos, G. Downey and L. Duponchel, *Food Chem.*, 2014, **148**, 124–130.
84 E. Borras, J. M. Amigo, F. V. D. Berg, R. Boque and O. Busto, *Food Chem.*, 2014, **153**, 15–19.
85 E. Capuano, G. V. D. Veer, R. Boerrigter-Eenling, A. Elgersma, J. Rademaker, A. Sterian and S. M. van Ruth, *Food Chem.*, 2014, **164**, 234–241.
86 H.-H. Gan, C. Soukoulis and I. Fisk, *Food Chem.*, 2014, **146**, 149–156.
87 S. A. Drivelos, K. Higgins, J. H. Kalivas, S. A. Haroutounian and C. A. Georgiou, *Food Chem.*, 2014, **165**, 316–322.
88 M. Silvestri, A. Elia, D. Bertelli, E. Salvatore, C. Durante, M. Li Vigni, A. Marchetti and M. Cocchi, *Chemom. Intell. Lab. Syst.*, 2014, **137**, 181–189.
89 M. R. D. Almeida, D. N. Correa, W. F. C. Rocha, F. J. O. Scafi and R. J. Poppi, *Microchem. J.*, 2013, **109**, 170–177.
90 *Encyclopedia of Spectroscopy and spectrometry*, ed. J. C. Lindom, G. E. Tranter and D. W. Koppennaal, Elsevier, Amsterdam, 3rd edn, 2017.
91 C. K. Muro, K. C. Doty, J. Bueno, L. Halamkova and I. K. Lednev, *Anal. Chem.*, 2015, **87**, 306–327.
92 J. Yang and Y.-y. Yang, *Pattern Recognit.*, 2003, **36**, 563–566.
93 H. Nocairi, E. M. Qannari, E. Vigneau and D. Bertrand, *Comput. Stat. Data Anal.*, 2005, **48**, 139–147.
94 D. V. Nguyen and D. M. Rocke, *Bioinformatics*, 2002, **18**, 39–50.
95 R. G. Brereton, *Chemom. Intell. Lab. Syst.*, 2015, **149**, 90–96.
96 E. K. Kemsley, *Chemom. Intell. Lab. Syst.*, 1996, **33**, 47–61.
97 M. Defernez and E. K. Kemsley, *TRAC, Trends Anal. Chem.*, 1997, **16**, 216–221.
98 N. Marigheto, E. Kemsley, M. Defernez and R. Wilson, *J. Am. Oil Chem. Soc.*, 1998, **75**, 987–992.
99 L. Tang, S. Peng, Y. Bi, P. Shan and X. Hu, *PLoS One*, 2014, **9**, e96944.
100 D. V. Nguyen and D. M. Rocke, *Bioinformatics*, 2002, **18**, 1216–1226.
101 P. Ciosek, Z. Brzozka, W. Wroblewski, E. Martinelli, C. D. Natal and A. D'AMico, *Talanta*, 2005, **67**, 590–596.
102 K. Kjedahl and R. Bro, *J. Chemom.*, 2010, **24**, 558–564.
103 P. Filzmoser, M. Gschwandtner and V. Todorov, *J. Chemom.*, 2012, **26**, 42–51.
104 R. G. Brereton, *TRAC, Trends Anal. Chem.*, 2006, **25**, 1103–1111.
105 N. F. Perez, J. Ferre and R. Boque, *Chemom. Intell. Lab. Syst.*, 2009, **95**, 122–128.
106 C. Botella, J. Ferre and R. Boque, *Talanta*, 2009, **80**, 321–328.
107 O. Galtier, O. Abbas, Y. L. Dreau, C. Rebufa, J. Kister, J. Artaud and N. Dupuy, *Vib. Spectrosc.*, 2011, **55**, 132–140.
108 D. Serrano-Lourido, J. Saurina, S. Hernandez-Cassou and A. Checa, *Food Chem.*, 2012, **135**, 1425–1431.
109 J. Engel, J. Gerretzen, J. E. Szymanska, J. J. Jansen, G. Downey, L. Blanchet and L. M. C. Budyens, *TRAC, Trends Anal. Chem.*, 2013, **50**, 96–106.
110 P. Lasch, *Chemom. Intell. Lab. Syst.*, 2012, **117**, 100–114.
111 L. C. Lee, C.-Y. Liong and A. A. Jemain, *AIP Conf. Proc.*, 2017, **1830**, 080008.
112 A. S. Rinnan, F. V. D. Berg and S. R. B. Engelsen, *TRAC, Trends Anal. Chem.*, 2009, **28**, 1201–1222.
113 T. Bocklitz, A. Walter, K. Hartmann, P. Rosch and J. Popp, *Anal. Chim. Acta*, 2011, **704**, 47–56.
114 A. R. Webb and K. D. Copsey, *Statistical Pattern Recognition*, Wiley, Chichester, 3rd edn, 2011.
115 I. Guyon and A. Elisseeff, *J. Mach. Learn. Res.*, 2003, **3**, 1157–1182.
116 H. Xie, J. Zhao, Q. Wang, Y. Sui, J. Wang, X. Yang, X. Zhang and C. Liang, *Sci. Rep.*, 2015, **5**, 10930.
117 M. Yin, S. Tang and M. Tong, *Anal. Methods*, 2016, **13**, 2794–2798.
118 N. S. Mirjankar, PhD Thesis, Institute of Chemical Technology, 2004.
119 J.-H. Cheng, D.-W. Sun and H. Pu, *Food Chem.*, 2016, **197**, 855–863.
120 G. Aliakbarzadeh, H. Parastar and H. Sereshti, *Chemom. Intell. Lab. Syst.*, 2016, **158**, 165–173.
121 T. Mehmood, K. H. Liland, L. Snipen and S. Saebo, *Chemom. Intell. Lab. Syst.*, 2012, **118**, 62–69.
122 O. Devos and L. Duponchel, *Chemom. Intell. Lab. Syst.*, 2011, **107**, 50–58.
123 C. M. Andersen and R. Bro, *J. Chemom.*, 2010, **24**, 728–737.
124 A. Issakson, M. Wallman, H. Goransson and M. G. Gustafsson, *Pattern Recognit. Lett.*, 2008, **29**, 1960–1965.
125 H. A. Martens and P. Dardenne, *Chemom. Intell. Lab. Syst.*, 1998, **44**, 99–121.
126 K. H. Esbensen and P. Geladi, *J. Chemom.*, 2010, **24**, 168–187.
127 S. Arlot, *Stat. Surveys*, 2010, **4**, 40–79.
128 R. Kohavi, in International Joint Conference on Artificial Intelligence (IJCAI), 1995, vol. 14, pp. 1157–1145.
129 L. Breiman and P. Spector, *Int. Stat. Rev.*, 1992, **60**, 291–319.
130 T. Hastie, R. Tibshirani and J. H. Friedman, in *The Elements of Statistical Learning, Data Mining, Inference and Prediction*, Springer, New York, 2009, ch. 7.10, pp. 214–217.
131 R. K. H. Galvao, M. C. U. Araujo, G. E. Jose, M. J. C. Pontes, E. C. Silva and T. C. B. Saldanha, *Talanta*, 2005, **67**, 736–740.
132 M. Daszykowski, B. Walczak and D. L. Massart, *Anal. Chim. Acta*, 2002, **468**, 91–103.
133 P. T. D. Goot, G. J. Postma, W. J. Melssen and L. M. C. Buydens, *Anal. Chim. Acta*, 1999, **392**, 67–75.

134 T. Borovicka, M. Jirina Jr., P. Kordik and M. Jirina, in *Advances in Data Mining Knowledge discovery and applications*, InTech, Croatia, 2012.
135 R. Wehrens, H. Putter and L. M. C. Buydens, *Chemom. Intell. Lab. Syst.*, 2000, **54**, 35–52.
136 P. Golland, F. Liang, S. Mukherjee and D. Panchenko, in *Learning Theory*, Springer, Berlin/Heidelberg, 2005, pp. 501–515.
137 http://wiki.eigenvector.com/index.php?title=Using_Cross-Validation.
138 R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137–148.
139 G. Quintas, N. Portillo, J. C. Garcia-Canaveras, J. V. Castell, A. Ferrer and A. Lahoz, *Metabolomics*, 2012, **8**, 86–98.
140 C. Rojas, R. Todeschini, D. Ballabio, A. Mauri, V. Consonni, P. Tripaldi and F. Grisoni, *Front. Chem.*, 2017, **5**, 53.
141 D. M. Hawkins and J. Kraker, *J. Chemom.*, 2010, **24**, 188–193.
142 A. M. Molinaro, R. Simon and R. M. Pfeiffer, *Bioinformatics*, 2005, **21**, 3301–3307.
143 P. Filzmoser, B. Liebmann and K. Varmuza, *J. Chemom.*, 2009, **23**, 160–171.
144 T. Hastie, R. Tibshirani and J. Friedman, The wrong and right way to do cross-validation, in *Elements of Statistical Learning, Data Mining, Inference, Prediction*, Springer, NY, 2009, pp. 245–247.
145 S. Chevallier, D. Bertrand, A. Kohler and P. Courcoux, *J. Chemom.*, 2006, **20**, 221–229.
146 M. Bylesjo, M. Rantalainen, O. Clorarec, J. K. Nicholson, E. Holmes and J. Trygg, *J. Chemom.*, 2006, **20**, 341–351.
147 G. James, D. Witten, T. Hastie and R. Tibshiranim, Assessing Model Accuracy, in *An introduction to statistical learning*, Springer, New York, 2013, pp. 29–36.
148 C. D. Brown and H. T. Davis, *Chemom. Intell. Lab. Syst.*, 2006, **80**, 24–38.
149 L. M. Reid, T. Woodcock, C. P. O'Donnell, J. D. Kelly and G. Downey, *Food Res. Int.*, 2005, **38**, 1109–1115.
150 S. J. Dixon, N. Heinrich, M. Holmboe M, M. L. Schaefer, R. R. Reed, J. Trevejo and R. G. Brereton, *Chemom. Intell. Lab. Syst.*, 2009, **99**, 111–120.
151 S. Smit, M. J. V. Breemen, H. C. Hoefsloot, A. K. Smilde, J. M. Aerts and C. G. D. Koster, *Anal. Chim. Acta*, 2007, **592**, 210–217.
152 B. Efron, *J. Am. Stat. Assoc.*, 1983, **78**, 316–331.
153 P. Refaeilzadeh, L. Tang and H. Liu, in *Encyclopedia of Database systems*, Springer, New York, 2009, pp. 532–538.
154 Q. S. Xu, Y. Z. Liang and Y.-P. Du, *J. Chemom.*, 2004, **19**, 112–120.
155 Q. S. Xu and Y. Z. Liang, *Chemom. Intell. Lab. Syst.*, 2001, **56**, 1–11.
156 B. Krakowska, D. Custers, E. Deconinck and M. Daszykowski, *Analyst*, 2015, **141**, 1060–1070.
157 A. J. Izenman, in *Modern Multivariate Statistical Techniques*, Springer, England, 2013, pp. 237–280.
158 R. G. Brereton, *J. Chemom.*, 2011, **25**, 225–246.
159 A. C. Lorena, L. F. O. Jacintho, M. F. Siqueira, R. De Giovanni, L. G. Lohmann, A. C. P. L. F. de Carvalho and M. Yamamoto, *Expert Syst. Appl.*, 2011, **38**, 5268–5275.
160 O. E. D. Noord, *Chemom. Intell. Lab. Syst.*, 1994, **23**, 65–70.
161 D. M. Hawkins, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 1–12.