Analytical and Bioanalytical Chemistry (2018) 410:5981–5992
https://doi.org/10.1007/s00216-018-1217-1

RESEARCH PAPER

Overoptimism in cross-validation when using partial least squares-discriminant analysis for omics data: a systematic study

Raquel Rodríguez-Pérez[1] · Luis Fernández[1,2] · Santiago Marco[1,2]

Received: 11 April 2018 / Revised: 13 June 2018 / Accepted: 21 June 2018 / Published online: 29 June 2018

Abstract

Advances in analytical instrumentation have provided the possibility of examining thousands of genes, peptides, or metabolites in parallel. However, the cost and time-consuming data acquisition process causes a generalized lack of samples. From a data analysis perspective, omics data are characterized by high dimensionality and small sample counts. In many scenarios, the analytical aim is to differentiate between two different conditions or classes combining an analytical method plus a tailored qualitative predictive model using available examples collected in a dataset. For this purpose, partial least squares-discriminant analysis (PLS-DA) is frequently employed in omics research. Recently, there has been growing concern about the uncritical use of this method, since it is prone to overfitting and may aggravate problems of false discoveries. In many applications involving a small number of subjects or samples, predictive model performance estimation is only based on cross-validation (CV) results with a strong preference for reporting results using leave one out (LOO). The combination of PLS-DA for high dimensionality data and small sample conditions, together with a weak validation methodology is a recipe for unreliable estimations of model performance. In this work, we present a systematic study about the impact of the dataset size, the dimensionality, and the CV technique used on PLS-DA overoptimism when performance estimation is done in cross-validation. Firstly, by using synthetic data generated from a same probability distribution and with assigned random binary labels, we have obtained a dataset where the true classification rate (CR) is 50%. As expected, our results confirm that internal validation provides overoptimistic estimations of the classification accuracy (i.e., overfitting). We have characterized the CR estimator in terms of bias and variance depending on the internal CV technique used and sample to dimensionality ratio. In small sample conditions, due to the large bias and variance of the estimator, the occurrence of extremely good CRs is common. We have found that overfitting peaks when the sample size in the training subset approaches the feature vector dimensionality minus one. In these conditions, the models are neither under- or overdetermined with a unique solution. This effect is particularly intense for LOO and peaks higher in small sample conditions. Overoptimism is decreased beyond this point where the abundance of noisy produces a regularization effect leading to less complex models. In terms of overfitting, our study ranks CV methods as follows: Bootstrap produces the most accurate estimator of the CR, followed by bootstrapped Latin partitions, random subsampling, K-Fold, and finally, the very popular LOO provides the worst results. Simulation results are further confirmed in real datasets from mass spectrometry and microarrays.

Keywords Metabolomics · Mass spectrometry · Microarrays · Chemometrics · Data analysis · Classification · Method validation

Introduction

Advances in analytical instrumentation have provided the possibility of examining thousands of genes, peptides, or metabolites in parallel. Technologies such as microarrays or mass spectrometry provide insight into system biology giving a large amount of complex biological data [1, 2]. However, the cost and time-consuming data acquisition process causes a generalized lack of samples. From a data analysis perspective, omics data are characterized by high dimensionality and

✉ Santiago Marco
smarco@ibecbarcelona.eu

[1] Signal and Information Processing for Sensing Systems, Institute for Bioengineering of Catalonia, The Barcelona Institute for Science and Technology, Baldiri Reixac 4-8, 08028 Barcelona, Spain

[2] Department of Electronics and Biomedical Engineering, University of Barcelona, Martí i Franqués 1, 08028 Barcelona, Spain

small sample counts [3]. Consequently, the "curse of dimensionality" [4] plays a key role and pattern recognition methods must be scrutinized in their ability to deal with small sample to dimensionality ratios [5, 6]. An additional factor in small sample conditions is the increased chances that the selected samples are not representative of the target population. In these conditions, the models become very dependent on the particular set of samples used for model building [7], and sometimes, they show poor generalization capabilities.

Model validation is an extremely important part in predictive model development. Beyond the numerical aspects, diverse levels of validation may serve to test for repeatability, reproducibility, instrumental shifts, or background (matrix) effects. Correct validation design including proper partition of the dataset taking into account data distribution in the input space, stratification issues, and replication of the future scenario of the model is essential to check the robustness of the models in the operational phase. These conceptual aspects of validation have been covered recently by Westad et al. [8] and Marco [9].

The simplest validation method is often known as "hold-out," and it refers to a simple splitting into training set and validation set. Even for this simple case, there are several alternatives for dataset splitting. Random sampling is the most popular, but it does not guarantee that samples in the borders of the set are within the training set. The Kennard-Stone algorithm (KS) [10] aims to sample the data space in a uniform manner maximizing the Euclidean distances between the selected samples. Recently, updates to the KS algorithm have been proposed to take into account the distribution of the dependent variable. An example is the SPXY splitting method by Galvao et al. [11]. This method was proposed in the context of building multivariate calibration models, that is, in a regression scenario. For classification problems, the distribution of the dependent variable is taken into account by trying to balance the partition so that the number of examples for the different classes is similar both in training and in validation.

Partial least squares-discriminant analysis (PLS-DA) is a common technique in multivariate analysis for classification or oriented dimensionality reduction. It has been the classifier of choice in multitude of applications in diverse fields [12–18]. Lately, PLS-DA algorithm has become a standard in omics research [19–28]. Some advantages behind the popularity of PLS-DA as classifier are the ability to cope with collinear and noisy variables, which is often the case in omics datasets [29], as well as possibility of results visualization by means of scores and loading plots [30, 31]. Additionally, variable importance can be interpreted from PLS-DA results [32, 33].

The propensity of PLS-DA to provide overoptimistic results (so called "overfitting"), and consequently poor generalization to samples outside the study, has been reported by several authors [31, 34, 35]. While there has been recent algorithmic proposals to reduce overfitting in PLS when noise among variables is not correlated [36], the usual approach is to optimize the number of latent variables (LV) such they show the best performance in validation.

A basic concept in this framework is to differentiate between internal and external validations. While in some publications, internal validation (IV) is also referred as validation and external validation (EV) as test or simply blind samples, we think this terminology (IV, EV) is less prone to confusion. IV refers to validation using those samples available for model building, while EV consists of fresh samples to test the performance of the model. The best practice is to use IV for model selection (e.g., optimizing the number of latent variables (LV), while performance estimation has to be performed using EV samples.

Beyond simple hold-out, for small sample counts, cross-validation (CV) methods aim at a more efficient use of the available data. In fact, many studies only report internal CV results skipping EV [30, 37]. However, it is well known that internal CV for performance estimation provides in general overoptimistic results, and an unbiased performance estimation should be done in an external validation set (also referred as "blind samples") [9, 38]. Despite this well-known fact, there is a very rich literature on different CV methods and their relative merits [39–41]. Over the years, many CV methods have been described in the literature, but a handful of them are the most popular in omics data analysis practice (see "Cross-validation").

On the other hand, methodologies have also been proposed to make optimum use of available samples together with EV methods. These methods are named double cross-validation [42]. A more general approach, it is known as cross-model validation (CMV) [43] and is often combined with jack-knifing model parameters [44]. Despite these well-known recommendations and methods, in omics research, simple CV is still the norm in most preliminary studies.

The scarce use of EV techniques in omics research is an issue that has been pointed out previously [45, 46]. Moreover, comparisons of CV methods for omics data [45–52] have already been published. Braga-Neto et al. compared linear discriminant analysis (LDA), which can be considered a particular case of PLS-DA [35], three-nearest neighbors (NN), and decision trees in an internal CV scheme. They concluded that CV had undesirable features, such as the presence of outliers in the accuracy estimation [53]. Fu et al. studied prediction errors with distinct CV strategies for random datasets of different sample sizes, but few cases of different dimensionality were considered [54]. Westerhuis compared single CV with CMV and advocated the use of permutation tests to compute the null hypothesis distribution of different figures of merit [34]. Finally, Varma et al. also highlighted the need of an independent set to estimate model's performance since

their CV and true error estimations differed in internal validation [55].

In this work, we explore the magnitude of overfitting of PLS-DA in internal CV. In particular, we will describe how overfitting depends on the chosen CV method, sample count, number of dimensions, and correlation among features. The study will be based initially in synthetic data, and then in real data from mass spectrometry and microarrays studies.

Methods

PLS-DA

In untargeted omics research, sample j is described by a feature vector $\mathbf{x}_j \in \mathbb{R}^D$ where D often takes large values. Feature vectors are acquired for n samples, providing a data matrix $\mathbf{X} \in \mathbb{R}^{nxD}$. Each sample has a phenotype or class label under study and thus can be described by a binary vector $\mathbf{y}_j \in \mathbb{R}^q$ where q is the number of categories (in the particular case of two classes it is enough to take $q = 1$). Then for n samples a categorical matrix $\mathbf{Y} \in \mathbb{R}^{nxq}$ might be defined as $\mathbf{Y} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$. In many omics studies, the aim is to build a suitable model that allows the accurate prediction of $\mathbf{y}_{new}$ from the measurement of $\mathbf{x}_{new}$.

PLS-DA can be understood as a partial least squares (PLS) regression between a set of predictors $\mathbf{X}$ and label responses $\mathbf{Y}$, with a binary outcome. PLS defines a new subspace of LV through an iterative process, considering a compromise between maximum variance in $\mathbf{X}$ and maximum correlation to $\mathbf{Y}$ [56, 57]. We focused on the algorithm PLS1, i.e., one response variable and multiple predictors. We employed the *plsr* function from the *pls* R package [58]. For PLSR models with one y-variable, no iterations are needed. The projection of the $\mathbf{X}$-matrix into the defined hyperplane is given by the $\mathbf{X}$-scores ($\mathbf{T}$), which are defined in Eq. (1).

$$\mathbf{T} = \mathbf{XW} \tag{1}$$

$$\mathbf{X} = \mathbf{TP}^{\mathbf{T}} + \mathbf{E} \tag{2}$$

$$\mathbf{Y} = \mathbf{TC}^{\mathbf{T}} + \mathbf{F} \tag{3}$$

$\mathbf{T}$ are result of the linear combination of the original variables with the weights $\mathbf{W}$, $\mathbf{T}$ model $\mathbf{X}$ (Eq. 2); when multiplied by the loadings $\mathbf{P}$, $\mathbf{X}$-scores are good summaries of $\mathbf{X}$ and the $\mathbf{X}$-residuals, $\mathbf{E}$, are small. On the other hand, $\mathbf{Y}$ can be predicted in terms of the $\mathbf{X}$-scores and the matrix $\mathbf{C}$ (Eq. 3). The $\mathbf{Y}$-residuals, $\mathbf{F}$, are the deviations between the observed and modeled responses. Finally, the relationship between $\mathbf{X}$ and $\mathbf{Y}$ that PLS specifies is given by Eq. (4), where $\mathbf{B}$ is a matrix with the PLS-regression coefficients (Eq. 5).

$$\mathbf{Y} = \mathbf{XB} + \mathbf{F} \tag{4}$$

$$\mathbf{B} = \mathbf{W}\left(\mathbf{P}^{\mathbf{T}}\mathbf{W}\right)^{-1}\mathbf{C}^{\mathbf{T}} \tag{5}$$

Cross-validation

Supervised algorithms require an estimation stage (also known as "training") with labeled data examples, and most classifiers also need tuning of hyper-parameters such as k in k-NN or LV number in the case of PLS-DA. We will refer as "internal validation" to the data split used for parameter optimization, and "external validation" to blind samples used to assess generalization capability or model's performance. Figure 1 shows the scheme of a three-way data split and the objectives of each data subset.

CV considers a number of iterations or folds with distinct training and test data partitions. For every fold, a model is built with the training set and tested for a range of hyper-parameter values. Finally, the selected hyper-parameter value is the one that provides the best average result along all the folds or partitions [59, 60]. We applied different CV strategies, namely: K-Fold, leave one out (LOO), random subsampling (RS), Bootstrap, and bootstrapped Latin partitions (BLP) for IV [1, 60, 61].

K-Fold

A dataset with n samples is split in k equal-sized partitions. The number of validation samples for each partition is n/k, and they must be in the validation set only once. We considered $k = 4$ that is 75–25% for training and test, respectively. Other choices of k are possible, but there is not a clear consensus on the preferred value for k.

Leave one out

LOO is the most extreme case of K-Fold. The training subset is composed by all samples but one, which is used for validation. The procedure is repeated $k = n$ times.

Random subsampling

In this strategy, the user can decide both the number of iterations and the percentage of training and validation subsets. Training samples are chosen randomly, and the rest are employed to validate, without resampling. Samples might be on the validation subset many times. We implemented a RS with 80 and 20 iterations, for the simulated and real datasets respectively, and a percentage of 75–25% for training and test.

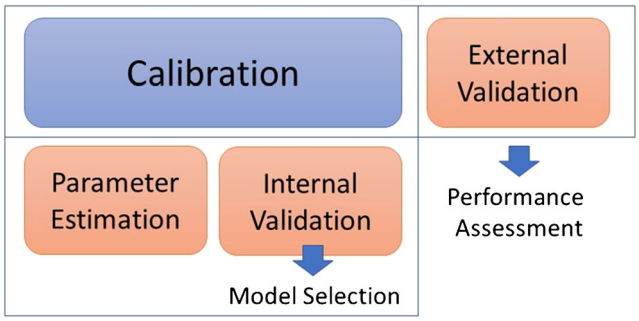

Fig. 1 Three-way data split. Full dataset is split in calibration subset and external validation subset, being the latter for predictive model performance assessment. Calibration subset is split in a training subset for parameter estimation and an internal validation subset for model selection

To ensure that class distribution among the training and validation samples is kept balanced, independent splitting of the two classes is performed and then merged. In this manner, we have a stratified partition.

Bootstrap

It is a resampling technique in which the user decides the number of iterations (about a hundred are recommended). In each iteration, *n* samples are chosen for training with replacement. The validation subset is formed by the rest of samples. Thus, training and test percentages depend on the replacement in every iteration. We performed bootstrapping with 80 iterations for the simulated and 20 and 100 iterations for real datasets [61–63]. Again, independent splitting of the two groups is performed and then merged in order to have a stratified partition.

Bootstrapped Latin partitions

In chemometrics, BLP have been considered a preferred approach to estimate the performance of predictive models and takes into account some typical characteristics of analytical instrumentation datasets [64–66]. A review of this technique has been recently published [67]. This method is a form of repeated K-Fold CV with some constraints. First, data are split so that replicates from the same chemical sample will not be contained in the prediction and training sets at the same time. In fact, the blind use of conventional CV methods in some works leads to the presence of very similar samples in training and validation and increase overoptimism. Second, the proportions of the number of samples for each class are automatically maintained between the validation set and training set. One of the advantages of BLP is that over CV repetitions, all samples are used for validation a single time. Selected examples for the use of this CV method can be found in Harrington et al. [68, 69] and Rearden et al. [71]. BLP combined with PLS and PLS-DA have been named super-PLS and super-PLS-DA, respectively, by Aloglu et al. [70]. A total of 80 and 20 iterations were performed for the synthetic and real sets, respectively.

Analysis

By overfitting, we understand the difference between the estimated classification accuracy in CV and the true accuracy. For simulated data the true accuracy is known by design (50% in this study), while in the case of real data true accuracy has to be estimated in external validation. We built simulated datasets with distinct samples to dimensionality ratios and feature correlation in order to evaluate PLS-DA overfitting under different conditions. Moreover, a mass spectrometry and a microarray dataset were used to ascertain whether simulation results were consistent with real data and how internal CV and external validation estimations differ when data distribution departs from normality or the expected accuracy is not 50%.

Simulated dataset

We created non-discriminative datasets using multivariate normal distributions obtained with *mvrnorm* function of the *MASS* R package. All samples were identically distributed irrespective of the class label, so that the theoretical discrimination power is null. In other words, for the synthetic dataset we know the true accuracy: 50% or random choice. Datasets were composed of Gaussian noise of mean $\mu = 0$ and covariance matrices with different correlation between features ($\sigma_{ii} = 1$, $\sigma_{ij} = \{0.00, 0.50, 0.90, 0.99\}$). For each dataset, two classes were arbitrarily defined by creating a random label binary vector with equal probability for both classes. PLS-DA overfitting was quantified in internal CV. LOO, K-Fold, RS, BLP, and Bootstrap methods were compared in terms of performance estimation. The number of LV was optimized according to the maximum average accuracy (classification rate (CR)) along the folds. Results of the best model are given as final performance estimation.

In order to study the influence of sample count *n* and dimensions *D*, we scanned both parameters in these ranges: $n \in (14,118)$ and $D \in (2100)$ both for training and internal CV. In all the considered cases, the two classes have the same number of instances and *n* is the sum. For each case of *n*, *D*, covariance matrix, and CV technique, we generated at least 1000 populations. For every population, the procedure to obtain the magnitude of overfitting was repeated. Estimator bias and its root mean square error (RMSE) were used as figures of merit (Eq. 6). According to the well know bias-variance decomposition [5], this error has two sources: Bias and Variance. Bias refers

to the expected difference between the true (CR_0) and the estimated ($\hat{CR}$) classification rates, whereas RMSE also takes the variance (Var) into account.

$$\mathrm{RMSE}(\hat{CR}) = \sqrt{E\left[(\hat{CR}-CR_0)^2\right]} = \sqrt{(\hat{CR}-CR_0)^2 + \mathrm{Var}(\hat{CR})} \quad (6)$$

It is important to remark that RMSE in this work does not refer to a mean square error between the numerical output of the PLS-DA model and the target label.

Microarray dataset

We employed a RNA microarray dataset which contains 295 samples from patients with *good* (110 samples) or *poor* (185 samples) (recurrence, distant metastasis or dead) prognosis after a mean follow-up of 6.7 years. In 2002, van't Veer et al. defined a gene expression profiling for breast cancer prognosis using this dataset). We used a 70-gene signature found by the authors to build PLS-DA models. This signature was previously used for assessing validity of CV for small-sample microarray classification [72, 73].

K-Fold, LOO, RS, BLP, and Bootstrap strategies were used to internally validate PLS-DA models. An unbiased estimation of the CR was obtained from external validation using 50 samples per class. We fixed the same number of data in training for all CV, which implies distinct total sample count. Models were balanced by considering the same number of samples of each class, and thus training and test sets had the same distribution (50–50%). To have a better estimation of the probability distribution of each estimator of the CR, we implemented a resampling strategy over the existing data: both internal and external validations were repeated 1000 times. From the obtained results, the estimator bias and variance were calculated for each internal validation method. Moreover, Wilcoxon tests were performed to assess the differences among validation methods and between IV and EV.

Mass spectrometer dataset

The public domain ARCENE dataset contains mass spectrometry data for patients with ovarian and prostate cancer, and control subjects. The data comes from two sources: National Cancer Institute (NCI) and Eastern Virginia Medical School (EVMS). Ovarian cancer data was obtained from NCI and prostate cancer from both sources. Spectra were preprocessed to minimize the disparity between data sources. The resultant training data was composed by 503 controls and 398 cancer samples, and 10,000 features (3000 of which were randomly permuted values to use the data for benchmark of feature selection methods). The employed version of the dataset was prepared by Isabelle Guyon and is available in the UCI Machine learning repository [74].

For this dataset, only LOO was used for validation. Candidate values for the sample count n and dimensions D were $n = \{14, 24, 54\}$ and $D = [2100]$. Again, n was set to even values so that every class had the same sample size. Feature selection was done by random selection, and for each case of sample and dimensionality (n, D), 500 independent trials were performed.

Results and discussion

Simulated dataset

Evaluation of overfitting using cross-validation

Since the occurrence of overfitting without validation is well known, many authors resort to internal validation for performance estimation. In this work, the number of LV has been optimized regarding the maximum average CR along all CV folds. The performance of the best model (i.e., with optimized complexity) is reported as the final accuracy estimation. Hence, the same validation data is used both for optimizing the model's parameters and estimating its performance.

Figure 2 includes accuracy estimations with K-Fold CV ($k = 4$) for increasing dimensionality and different sample sizes. The figure shows a significant overfitting that peaks when the dimensionality matches the number of samples in training minus one. The importance of this overfitting increases in small sample conditions. Beyond the peak and contrary with intuition, overfitting decreases as the number of dimensions becomes larger than the number of samples. This behavior cannot be observed when executing few times

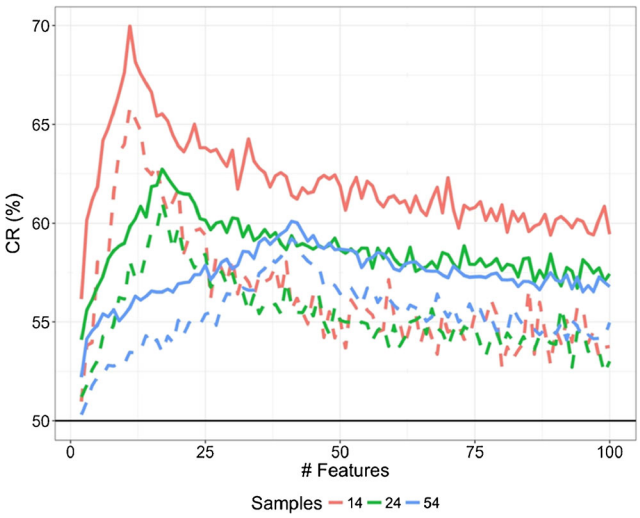

Fig. 2 CR estimation in fourfold CV for simulated data. Cases with 0.9 correlation (solid line) and without correlation (dashed line). Mean after 1000 runs for each D ($n = 50$)

each condition (n, D), but when averaging thousands of repetitions. Furthermore, it is closely related with the complexity of the PLS-DA models. The frequency distribution of the optimum LV number follows the same tendency. Therefore, the sample to dimensionality ratio of the training data strongly affects to the complexity of the optimal model, which is maximum when the number of training samples approaches the number of features (i.e., determined system). Hence, at the overfitting peak, the maximum number of LV is frequently selected.

On the other hand, Fig. 2 shows that data with covariance matrix $\Sigma = (\sigma_{ii} = 1, \sigma_{ij} = 0.9)$ lead to more overoptimistic results than data without correlated features. Consequently, more multicollinearity among the independent variables caused more overfitting. At any rate, the qualitative behavior is the same and the magnitude of overfitting is large in both cases. We would like to remark that at the peak, overfitting can reach a mean of 70 or 67% in CR for the cases with 0.9 and 0.0 correlation, respectively. Since omics data usually contains correlated features, this highlights the need to examine results with a critical eye and even the adequacy of using independent subsets of variables.

For a more comprehensive comparison, we should consider not only the bias of the estimator but also the variance. Thus, we computed the spread of the estimated accuracy with fourfold CV. Figure 3 shows the distributions for $n = 18$ and 118, approximately in the peak of overfitting, i.e., $D = 13$ and 89 features ($D = n_{\text{training}} - 1$). Both CR estimations are biased and have a certain variance, but a higher sample count provides lower bias and variance. Further, our results indicate that even with a true CR of 50%, internal K-Fold CV may give classification rates over 90% when the sample size is small (e.g., $n = 18$). Whether K-Fold CV is executed once, estimation might provide an extremely good or bad result only by chance. Consequently, this highlights the need of permutations test in order to know accuracy distribution of a random classification and assess the significance of the results compared with chance.

Distinct CV strategies were compared for many sample-to-dimensionality ratios in synthetic data. Figure 4 shows the mean CR after repeating a thousand times the population generation and the internal CV estimations for $n_{\text{total}} = 50$ and each case of D. This figure suggests that overfitting follows the same qualitative behavior for every CV method, i.e., overfitting peaks when the dimensionality approaches the number of training samples, except for Bootstrap, which shows a more flattened curve. Training sample size is 49 for LOO and 38 for K-Fold and RS, but Bootstrap does not define a single value of samples in test for each iteration. However, it is known that in average training is a 63.2% of the data, so approximately 32 different samples for training and 18 for test. Precisely for the latter reason, Bootstrap is the only method which does not show a sharp peak of overfitting. Our

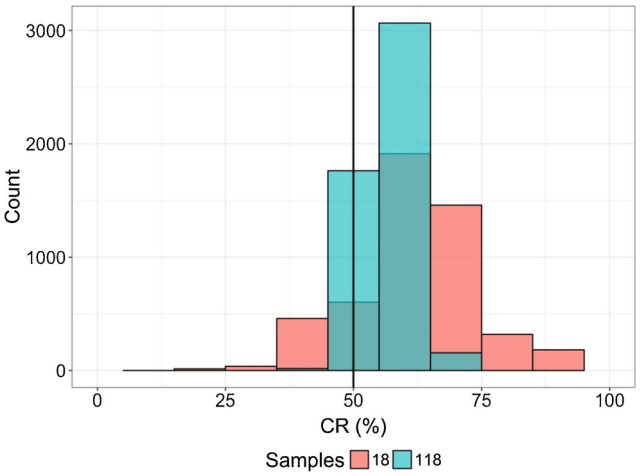

Fig. 3 CR distributions in fourfold CV for simulated data $N(0, 1)$. Cases with $(n, D) = (18, 13)$ and $(n, D) = (118, 89)$. Mean after 5000 runs for each pair (n,D)

results show that the magnitude of bias changes among CV strategies. Specifically, LOO and K-Fold seem to be the most biased, while RS, BLP, and Bootstrap give more accurate estimations. These latter approaches have a large number of resampling iterations and seem to make better use of the available data. When the number of features approaches to sample size in training, LOO produces higher overfitting than K-Fold. On the contrary, K-Fold introduces more bias than LOO far from the peak. Moreover, K-Fold gives higher overoptimistic estimations than RS or BLP. RS and BLP give essentially the same results and closely follow Bootstrap approach. Precisely, Bootstrap gives the least biased estimation, whose maximum bias is of only 3% for the simulated dataset. Current results indicate that for high dimensionality scenarios, RS and BLP are less biased than Bootstrap, while when the number of features decreases, this trend reverses.

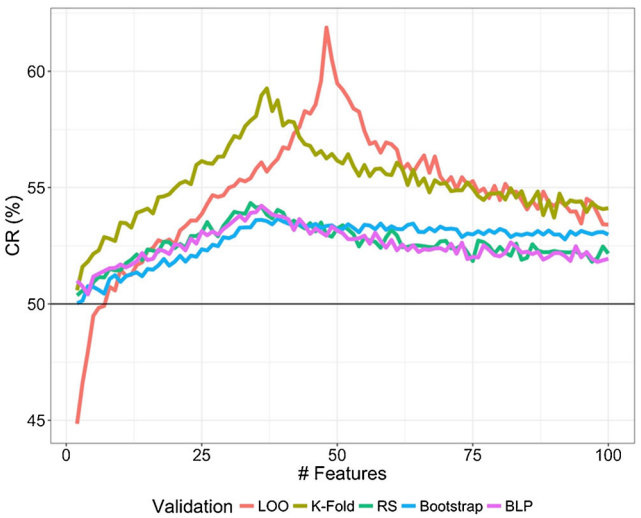

Fig. 4 Mean CR of PLS-DA models with five cases of internal CV: LOO, K-Fold, RS, Bootstrap, and BLP. Mean after 1000 runs for each D with simulated data $N(0,,1)$ and $n = 50$

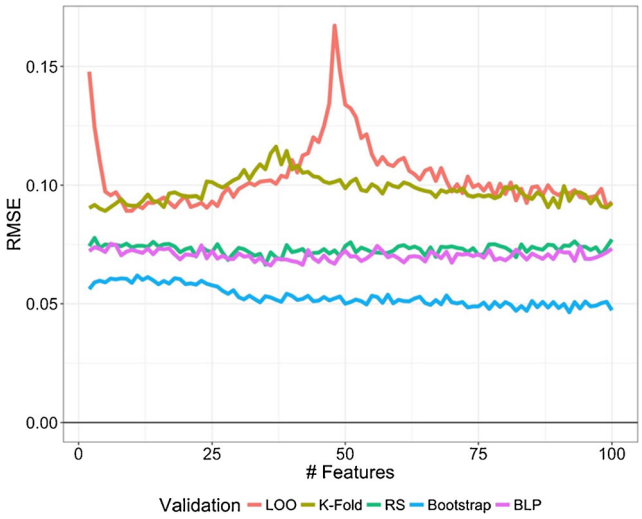

Fig. 5 Mean RMSE of PLS-DA models with five cases of internal CV: LOO, K-Fold, RS, Bootstrap, and BLP. Mean after 1000 runs for each D with simulated data $N(0, 1)$ and $n = 50$

In each CV approach, peaks are located at different dimensionality since training subsets have different sizes for a given number of total samples. Systematical computations of CR along different conditions suggest that a number of training samples similar to the dataset dimensions is a condition of analysis which should be avoided, since it appears to be the worst scenario in terms of overfitting.

We also evaluated RMSE of the CR estimator in order to consider both bias and variance in the overfitting assessment. After repeating a thousand generations and evaluations of datasets with distinct (n, D) pairs, we represented the average RMSE (Fig. 5). These models were optimized and evaluated with the five presented methods of CV. The risk of obtaining overoptimistic was shown to depend on the sample to dimensionality ratio and the CV strategy employed, as previously hypothesized. A trade-off between bias and variance was observed, causing that conditions of minimum bias corresponded to maximum variance, and conversely. This evaluation let us clearly rank the CV methods in terms of their ability to provide accurate estimations of the CR. In this sense, these results advocate the use of Bootstrap, followed by BLP, RS, K-Fold, and finally LOO. Please observe the important peak of LOO in terms of RMSE of the estimator. To put these results into context, the reader should remember that in many occasions omics datasets are characterized by small sample counts and in those conditions, LOO is the preferred validation technique by many authors [75–80]. Finally, while Bootstrap seems to be the best CV procedure according to RMSE criterion, it is important to realize that it will always introduce some overoptimism in the estimation of the CR.

Microarray dataset

We utilized a RNA microarray dataset containing 70 gene-expression features to build PLS-DA binary classification models for breast cancer prognosis. For this dataset, the expected CR is not 50%. We set sample size to the maximum overfitting condition according to the conclusions derived from the previous experiments, so we established that sample count in training minus one equaled the number of features. Since the number of genes is 70, we set the number of samples in training to be 71 in every case, which implies a different total number of examples for every CV. This condition of sample size allowed to compare CV techniques in a scenario of significant overfitting, particularly it would correspond to compare them in the peak of overfitting.

To evaluate the overfitting introduced by the CV method, we have to resort to EV as an unbiased estimation of the CR. The main problem here is since the EV set will have a finite number of samples, the estimated CR will have some unavoidable variance. The number of EV samples was always 50 for each class, which corresponded to the minimum number of remaining samples among all the CV cases, to have comparable test results.

Table 1 shows the results of CR in internal and external validation, as well as the RMSE of the estimator. We can observe that the ascending rank of RMSE, which considers both bias and variance, is: Bootstrap, BLP, RS, K-Fold, and LOO, which coincides with the one obtained with simulated datasets. Comparing K-Fold with LOO, we can observe that a larger k leads to a variance increment. K-Fold, RS, and BLP have the same number of samples in training, but in RS and BLP, which have more folds or iterations than K-Fold, the bias diminishes.

On the other hand, it is interesting to highlight that different CV techniques produce PLS-DA models with diverse accuracy. The table shows that Bootstrap provides the best CR in EV, whereas LOO leads to the worst case of predictive accuracy.

Table 1 Comparison of CV strategies in terms of CR in IV and EV and RMSE of the estimator in IV

CV	LOO	K-Fold	RS	BLP	Bootstrap
Internal CR (%)	69.1 ± 7.5	66.9 ± 4.1	66.1 ± 4.3	66.1 ± 3.9	65.5 ± 3.9
External CR (%)	60.1 ± 7.0	62.5 ± 5.1	63.8 ± 4.7	64.1 ± 4.5	64.6 ± 4.4
RMSE (%)	11.7	6.0	4.9	4.4	4.0

Mean and standard deviation of CRs are reported

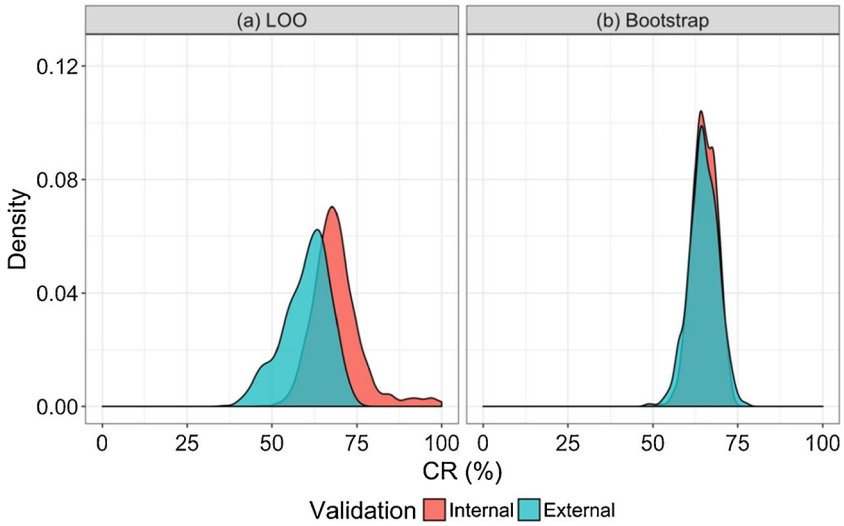

Fig. 6 CR distributions of internal CV and EV hold out for real microarray data. **a** Internal LOO CV and **b** internal Bootstrap

Thus, Bootstrap appears to be the best approach to optimize model complexity not only in terms of better performance estimation but also in terms of producing the more accurate model. From the inspection of this table, we would like to remark that models were the overfitting is bigger typically result in poorer results in EV. In other words, the selection of the CV method is key, not only to have small bias but also to obtain the most accurate model. The results of the 1000 trials were used to test whether IV and EV results were statistically different for all the CV methods. Bootstrap with 100 iterations was the only case in which the null hypothesis of equality between internal and external CR distributions could not be rejected. Moreover, when comparing the estimations between the different methods, all of them were statistically different except RS and BLP (p values of 0.762 and 0.227 for IV and EV, respectively). External CRs of BLP and Bootstrap also had a p value close to the significance level of 5% (p value = 0.041), but the rest of the p values were below 0.01, indicating that under these conditions the magnitude of overfitting of the methods is statistically different.

Figure 6 shows the accuracy distributions of internal CV and EV, for the cases of LOO (Fig. 6a) and Bootstrap (Fig. 6b). These plots clearly depict the extent of the overfitting in a case where real discrimination between both classes does exist. In the case of LOO (Fig. 6a), we can clearly see how the distribution is shifted to the right and has a tail towards very high accuracies. This effect is much less evident in the case of Bootstrap, as Fig. 6b shows. In Fig. 6b, Bootstrap was computed for 100 iterations since it is a value typically encountered.

Together with a larger bias, LOO CV yields to an estimation with more variance than Bootstrap. In fact, the peaks of the density distributions for Bootstrap and internal and external distributions are almost fully overlapped. The latter feature is highly desired because it means that internal CV results almost provide the same statistics as EV.

In summary, this study with real microarray data confirms the dependence of the overfitting with respect to the CV technique implemented. It confirms that LOO can be considered a weak validation practice while Bootstrap provides more accurate performance estimations with less bias and variance.

Mass spectrometry data

First, we report the obtained results for randomly selected features. Due to the complexity of the dataset, random-selected features have a null discriminant power when this is estimated in EV. However, in IV again, we have a clear over-optimistic bias as expected.

The mean classification rate when computed over the varied selection of features are shown in Fig. 7. As it was already observed for simulated data we can see how we obtain a peak when the number of features approaches the number of

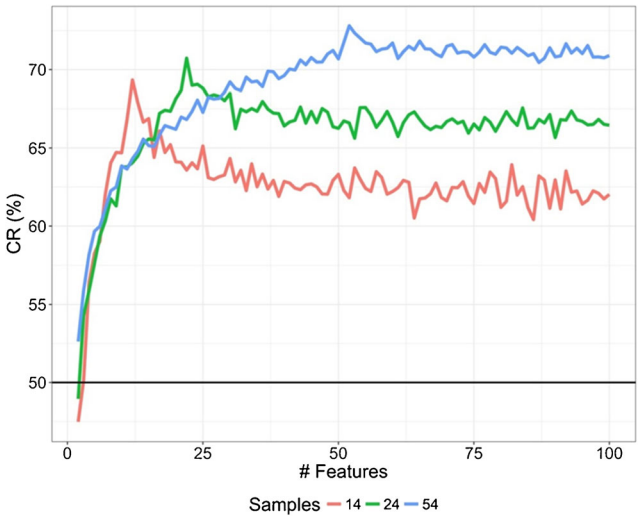

Fig. 7 CR estimation by LOO for the ARCENE dataset with random feature selection. The mean CR after 500 trials for each D is shown for n = 14, 24, and 54 samples

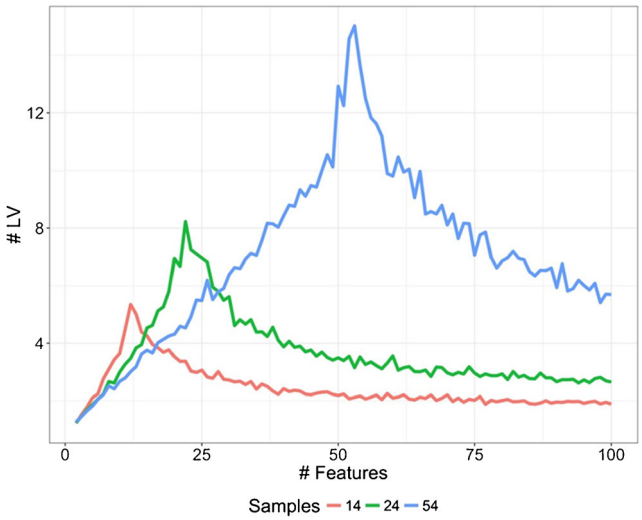

Fig. 8 Optimum number of LV selected by LOO in the ARCENE dataset. The mean number of LV after 500 trials for each D is shown for $n = 14$, 24, and 54 samples

samples. Before that peak, overoptimism increases with the number of features. Beyond the peak, the overoptimism decreases slightly and then saturates. In this region, we can observe a counter-intuitive behavior since, fixed the number of features, we have more overoptimism when we have more samples in the dataset. Similarly, we obtain less overoptimism increasing the number of features, for a given number of data samples.

This behavior is obviously related to the average complexity of the models. In Fig. 8, we observe that the model complexity (in terms of number of LV) also peaks when the number of features equals the number of samples.

Beyond the evolution of the bias, we can observe how the full probability density function for the estimator behaves. In the case of only 14 samples, Fig. 9 shows the probability distribution for the estimator of the CR in internal and external validation. Figure 9a corresponds to three features, while Fig. 9b corresponds to 13 features. In both cases, the estimator in LOO is biased and has a much larger variance than the estimator in EV. Figure 9b clearly shows how bias increases when the number of features approaches the number of samples.

Conclusions

PLS-DA is a preferred predictive model for the analysis of omics data, particularly in the case of metabolomics. Metabolomics data are usually characterized by large dimensionalities and small sample conditions. In these conditions, PLS-DA models have a very strong propensity to overfit to training data. Despite external validation being the recommended practice, many researchers still prefer simple cross-validation in most studies with small datasets.

The need of stronger validation practices such as CMV and permutation test has already been highlighted in the prior literature. The main message of this work is a full characterization of the impact of the sample size, the dataset dimensionality, and the CV method on the overfitting.

Therefore, we have shown a strong dependence of PLS-DA with the sample to dimensionality ratio. For the first time, a full scan of the impact of the dataset size, dimensionality ratio, and the CV technique is done. In extreme cases, for small datasets and using LOO, mean overfit may exceed 20% in a case where there is of no discriminant information. We have observed that for a given number of data samples, increasing the dimensionality leads to more complex models that are obviously easier to overfit. However, the maximum number of LV is limited to the number of data minus one.

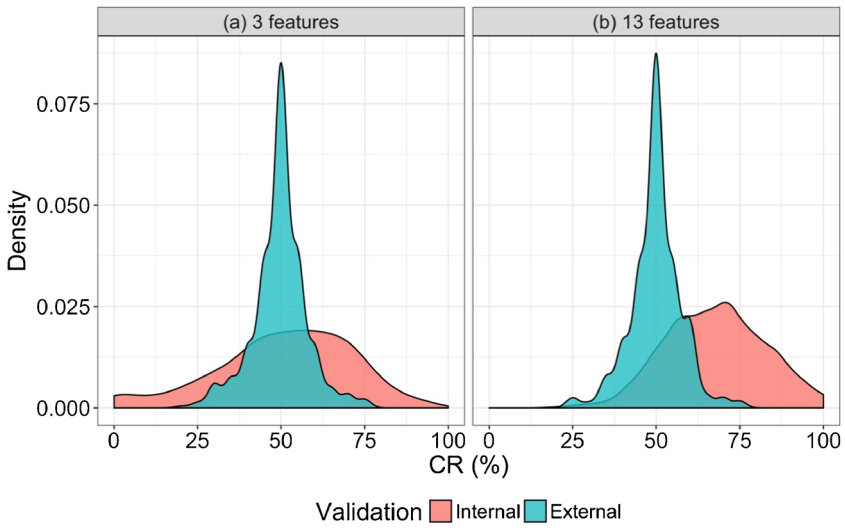

Fig. 9 Distribution of the CR in internal CV (LOO) and EV after 500 repetitions for the case 14 samples and **a** 3 and **b** 13 features

From that point on, increasing feature vector dimensionality does not allow for more complex models, but instead the additional features are a source of noise and provide a regularization effect in the complexity of the models, leading to simpler models with less overfit. We have shown that the PLS-DA overfitting in CV peaks when the sample size in the training set approaches to the number of dimensions of the dataset. In addition, it decreases far from the peak even if the number of dimensions is much larger than the number of samples. This should be chiefly considered in dimensionality reduction prior to PLS-DA modeling, since training samples matching the number of dimensions appears to be a scenario to avoid. As it has been suggested previously, permutation tests help in determining if the obtained results are likely to be obtained by chance.

Among all the CV techniques, Bootstrap provides the most accurate estimator in terms of RMSE, followed by RS and BLP. In fact, for the microarray case under study, the internal and external validation estimations were almost equal in the case of Bootstrap. Resampling validation techniques provide the most efficient use of the available data. Instead, LOO appeared to provide estimations with large variance and also with more bias than the other CV strategies. This result is important in omics research due to the popularity of LOO in the small sample count datasets typically encountered. Additionally, the models obtained by LOO show a degraded performance when evaluated in external validation when compared with other CV techniques.

This work highlights the need of strong validation methodologies to be used in conjunction with PLS-DA, since the uncritical use of these techniques may lead to overoptimistic results and contribute to the irreproducibility problem in omics research. A rigorous validation strategy is key to avoid overfitting. Hence, we strongly encourage the use of external validation to obtain an unbiased estimation of model's predictive performance (e.g., double CV or CMV) and permutation tests to evaluate whether the obtained CR is statistically significant.

Authors' contributions RR wrote the software, analyzed the data, and prepared the figures and text. LF supervised the code of RR and provided useful insights. SM conceived the study and supervised the work. RR and SM authors contributed to writing the manuscript. All authors read and approved the final manuscript.

Funding information This work was partially funded by the Spanish MINECO program, under grants TEC2011-26143 (SMART-IMS) and TEC2014-59229-R (SIGVOL). The Signal and Information Processing for Sensor Systems group is a consolidated Grup de Recerca de la Generalitat de Catalunya and has support from the Departament d'Universitats, Recerca i Societat de la Informació de la Generalitat de Catalunya (expedient 2017 SGR 1721). This work has received support from the Comissionat per a Universitats i Recerca del DIUE de la Generalitat de Catalunya and the European Social Fund (ESF). Additional financial support has been provided by the Institut de Bioenginyeria de Catalunya (IBEC). IBEC is a member of the CERCA Programme/Generalitat de Catalunya.

Compliance with ethical standards

Ethics approval and consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material The microarray dataset analyzed during the current study is publicly available at http://ccb.nki.nl/data/.

Competing interests The authors declare that they have no competing interests.

References

1. Santana R, Galdiano J, Pérez A, Bielza C, Larrañaga P, Calvo B, et al. Machine learning in bioinformatics machine learning in bioinformatics. Brief Bioinform. 2006;7:1–16. https://doi.org/10.1093/bib/bbk007.
2. Kulasingam V, Diamandis EP. Strategies for discovering novel cancer biomarkers through utilization of emerging technologies. Nat Clin Pract Oncol. 2008;5:588–99. https://doi.org/10.1038/ncponc1187.
3. Vinaixa M, Samino S, Saez I, Duran J, Guinovart JJ, Yanes O. A guideline to univariate statistical analysis for LC/MS-based untargeted metabolomics-derived data. Metabolites. 2012;2:775–95. https://doi.org/10.3390/metabo2040775.
4. Bellman R. Adaptive control processes—a guided tour. Z Angew Math Mech. 1962;42:364–5.
5. Bishop CM. Pattern recognition and machine learning. Heidelberg: Springer-Verlag Berlin; 2006.
6. Ghosh D, Poisson LM. "Omics" data and levels of evidence for biomarker discovery. Genomics. 2009;93:13–6. https://doi.org/10.1016/j.ygeno.2008.07.006.
7. Rubingh CM, Bijlsma S, Derks EPP, Bobeldijk I, Verheij ER, Kochhar S, et al. Assessing the performance of statistical validation tools for megavariate metabolomics data. Metabolomics. 2006;2:53–61. https://doi.org/10.1007/s11306-006-0022-6.
8. Westad F, Marini F. Validation of chemometric models—a tutorial. Anal Chim Acta. 2015;893:14–24. https://doi.org/10.1016/j.aca.2015.06.056.
9. Marco S. The need for external validation in machine olfaction: emphasis on health-related applications chemosensors and chemoreception. Anal Bioanal Chem. 2014;406:3941–56. https://doi.org/10.1007/s00216-014-7807-7.
10. Kennard RW, Stone LA. Computer aided design of experiments. Technometrics. 1969;11:137–48. https://doi.org/10.1080/00401706.1969.10490666.
11. Galvão RKH, Araujo MCU, José GE, Pontes MJC, Silva EC, Saldanha TCB. A method for calibration and validation subset partitioning. Talanta. 2005;67:736–40. https://doi.org/10.1016/j.talanta.2005.03.025.
12. Barker M, Rayens W. Partial least squares for discrimination. J Chemom. 2003;17:166–73. https://doi.org/10.1002/cem.785.
13. Chevallier S, Bertrand D, Kohler A, Courcoux P. Application of PLS-DA in multivariate image analysis. J Chemom. 2006;20:221–9. https://doi.org/10.1002/cem.994.

14. Sirven J-B, Sallé B, Mauchien P, Lacour J-L, Maurice S, Manhès G. Feasibility study of rock identification at the surface of Mars by remote laser-induced breakdown spectroscopy and three chemometric methods. J Anal At Spectrom. 2007;22:1471. https://doi.org/10.1039/b704868h.
15. Ciosek P, Wróblewski W. Miniaturized electronic tongue with an integrated reference microelectrode for the recognition of milk samples. Talanta. 2008;76:548–56. https://doi.org/10.1016/j.talanta.2008.03.051.
16. Ivorra E, Girón J, Sánchez AJ, Verdú S, Barat JM, Grau R. Detection of expired vacuum-packed smoked salmon based on PLS-DA method using hyperspectral images. J Food Eng. 2013;117:342–9. https://doi.org/10.1016/j.jfoodeng.2013.02.022.
17. Bassbasi M, De Luca M, Ioele G, Oussama A, Ragno G. Prediction of the geographical origin of butters by partial least square discriminant analysis (PLS-DA) applied to infrared spectroscopy (FTIR) data. J Food Compos Anal. 2014;33:210–5. https://doi.org/10.1016/j.jfca.2013.11.010.
18. Lo Y-L, Pan W-H, Hsu W-L, Chien Y-C, Chen J-Y, Hsu M-M, et al. Partial least square discriminant analysis discovered a dietary pattern inversely associated with nasopharyngeal carcinoma risk. PLoS One. 2016. https://doi.org/10.1371/journal.pone.0155892.
19. Pérez-Enciso M, Tenenhaus M. Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach. Hum Genet. 2003;112:581–92. https://doi.org/10.1007/s00439-003-0921-9.
20. Boulesteix AL, Strimmer K. Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. Brief Bioinform. 2007;8:32–44. https://doi.org/10.1093/bib/bbl016.
21. Izquierdo-García JL, Rodríguez I, Kyriazis A, Villa P, Barreiro P, Desco M, et al. A novel R-package graphic user interface for the analysis of metabonomic profiles. BMC Bioinformatics. 2009;10. https://doi.org/10.1186/1471-2105-10-363.
22. Biswas A, Mynampati KC, Umashankar S, Reuben S, Parab G, Rao R, et al. Metdat: a modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation. Bioinformatics. 2010;26:2639–40. https://doi.org/10.1093/bioinformatics/btq436.
23. Smolinska A, Blanchet L, Buydens LMC, Wijmenga SS. NMR and pattern recognition methods in metabolomics: from data acquisition to biomarker discovery: a review. Anal Chim Acta. 2012;750:82–97. https://doi.org/10.1016/j.aca.2012.05.049.
24. Sugimoto M, Kawakami M, Robert M, Soga T, Tomita M. Bioinformatics tools for mass spectroscopy-based metabolomic data processing and analysis. Curr Bioinforma. 2012;7:96–108. https://doi.org/10.2174/157489312799304431.
25. Cauchi M, Fowler DP, Walton C, Turner C, Jia W, Whitehead RN, et al. Application of gas chromatography mass spectrometry (GC-MS) in conjunction with multivariate classification for the diagnosis of gastrointestinal diseases. Metabolomics. 2014;10:1113–20.
26. Bro R, Kamstrup-Nielsen MH, Engelsen SB, Savorani F, Rasmussen MA, Hansen L, et al. Forecasting individual breast cancer risk using plasma metabolomics and biocontours. Metabolomics. 2015;11:1376–80. https://doi.org/10.1007/s11306-015-0793-8.
27. Garreta-Lara E, Campos B, Barata C, Lacorte S, Tauler R. Metabolic profiling of *Daphnia magna* exposed to environmental stressors by GC–MS and chemometric tools. Metabolomics. 2016;12. https://doi.org/10.1007/s11306-016-1021-x.
28. Fang J, Wang W, Sun S, Wang Y, Li Q, Lu X, et al. Metabolomics study of renal fibrosis and intervention effects of total aglycone extracts of Scutellaria baicalensis in unilateral ureteral obstruction rats. J Ethnopharmacol. 2016;192:20–9. https://doi.org/10.1016/j.jep.2016.06.014.
29. Lämmerhofer M, Weckwerth W. Metabolomics in practice successful strategies to generate and analyze metabolic data. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA; 2013.
30. Broadhurst DI, Kell DB. Statistical strategies for avoiding false discoveries in metabolomics and related experiments. Metabolomics. 2006;2:171–96. https://doi.org/10.1007/s11306-006-0037-z.
31. Gromski PS, Muhamadali H, Ellis DI, Xu Y, Correa E, Turner ML, et al. A tutorial review: metabolomics and partial least squares-discriminant analysis - a marriage of convenience or a shotgun wedding. Anal Chim Acta. 2015;879:10–23. https://doi.org/10.1016/j.aca.2015.02.012.
32. Eriksson L, Johansson E, Kettaneh-Wold N, Wold S. Introduction to multi-and megavariate data analysis using projection methods (PCA & PLS). Umea: Umetrics AB; 1999.
33. Mehmood T, Liland KH, Snipen L, Saebø S. A review of variable selection methods in partial least squares regression. Chemom Intell Lab Syst. 2012;118:62–9. https://doi.org/10.1016/j.chemolab.2012.07.010.
34. Westerhuis JA, Hoefsloot HCJ, Smit S, Vis DJ, Smilde AK, Velzen EJJ, et al. Assessment of PLSDA cross validation. Metabolomics. 2008;4:81–9. https://doi.org/10.1007/s11306-007-0099-6.
35. Brereton RG, Lloyd GR. Partial least squares discriminant analysis: taking the magic away. J Chemom. 2014;28:213–25. https://doi.org/10.1002/cem.2609.
36. Sousa PF, Åberg KM. Can we beat overfitting?—a closer look at Cloarec's PLS algorithm. J Chemom. 2018:e3002. https://doi.org/10.1002/cem.3002.
37. Agne K, Alexander HJ, Marcis L, Juozas K, Hossam H, Hermann B. Detection of cancer through exhaled breath: a systematic review. Oncotarget. 2015;6. https://doi.org/10.18632/oncotarget.5938.
38. Steyerberg EW, Bleekerb SE, Moll HA, Grobbee DE, Moons KGM. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. J Clin Epidemiol. 2003;56:441–7. https://doi.org/10.1016/S0895-4356(03)00047-7.
39. Kim J-H. Estimating classification error rate: repeated cross-validation, repeated hold-out and Bootstrap. Comput Stat Data Anal. 2009;53:3735–45. https://doi.org/10.1016/J.CSDA.2009.04.009.
40. Jiang G, Wang W. Error estimation based on variance analysis of k-fold cross-validation. Pattern Recogn. 2017;69:94–106. https://doi.org/10.1016/j.patcog.2017.03.025.
41. Wong TT. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. Pattern Recogn. 2015;48:2839–46. https://doi.org/10.1016/j.patcog.2015.03.009.
42. Filzmoser P, Liebmann B, Varmuza K. Repeated double cross validation. J Chemom. 2009;23:160–71. https://doi.org/10.1002/cem.1225.
43. Anderssen E, Dyrstad K, Westad F, Martens H. Reducing overoptimism in variable selection by cross-model validation. Chemom Intell Lab Syst. 2006;84:69–74. https://doi.org/10.1016/J.CHEMOLAB.2006.04.021.
44. Martens H, Martens M. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). Food Qual Prefer. 2000;11:5–16. https://doi.org/10.1016/S0950-3293(99)00039-7.
45. Kjeldahl K, Bro R. Some common misunderstanding in chemometrics. J Chemom. 2010;24:558–64.
46. Xia J, Broadhurst DI, Wilson M, Wishart DS. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. Metabolomics. 2013;9:280–99. https://doi.org/10.1007/s11306-012-0482-9.
47. Kohavi R (2016) A study of cross-validation and Bootstrap for accuracy estimation and model selection. IJCAI'95 Proceedings of the 14th International Joint Conference on Artificial Intelligence 2:1137–1143.

48. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005;21:3301–7. https://doi.org/10.1093/bioinformatics/bti499.
49. Wood I, Visscher PM, Mengersen KL. Classification based upon gene expression data: bias and precision of error rates. Bioinformatics. 2007;23:1363–70. https://doi.org/10.1093/bioinformatics/btm117.
50. Boulesteix AL, Strobl C. Optimal classifier selection and negative bias in error rate estimation: an empirical study on high-dimensional prediction. BMC Med Res Methodol. 2009;9. https://doi.org/10.1186/1471-2288-9-85.
51. Szymańska E, Saccenti E, Smilde AK, Westerhuis JA. Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies. Metabolomics. 2012;8:3–16. https://doi.org/10.1007/s11306-011-0330-3.
52. Triba MN, Le Moyec L, Amathieu R, Goossens C, Bouchemal N, Nahon P, et al. PLS/OPLS models in metabolomics: the impact of permutation of dataset rows on the K-fold cross-validation quality parameters. Mol BioSyst. 2015;11:13–9. https://doi.org/10.1039/C4MB00414K.
53. Braga-Neto UM, Dougherty ER. Is cross-validation valid for small-sample microarray classification? Bioinformatics. 2004;20:374–80. https://doi.org/10.1093/bioinformatics/btg419.
54. Fu WJ, Carroll RJ, Wang S. Estimating misclassification error with small samples via Bootstrap cross-validation. Bioinformatics. 2005;21:1979–86. https://doi.org/10.1093/bioinformatics/bti294.
55. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;10:91. https://doi.org/10.1186/1471-2105-7-91.
56. Phatak A, De Jong S. The geometry of partial least squares. J Chemom. 1997;11:311–38. https://doi.org/10.1002/(SICI)1099-128X(199707)11:4<311::AID-CEM478>3.0.CO;2-4.
57. Wold SSM, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst. 2001;58:109–30.
58. Mevik B-HBHB, Wehrens R. The pls package: principal component and partial least squares regression in R. J Stat Softw. 2007;2007:18.
59. Stone M. Cross-validatory choice and assessment of statistical predictions. J R Stat Soc. 1974;36:111–47. https://doi.org/10.2307/2984809.
60. Burman P. A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning testing methods. Biometrika. 1989;76:503–14.
61. Efron B, Tibshirani R. Estimating the error rate of a prediction rule. J Am Stat Assoc. 1983;78:316–31. https://doi.org/10.1080/01621459.1983.10477973.
62. Efron B, Tibshirani R. Improvements on cross-validation: the 632+ Bootstrap method. J Am Stat Assoc. 1997;92:548–60.
63. Brereton R. Chemometrics for pattern recognition. Chichester: Wiley; 2009.
64. de Boves HP. Statistical validation of classification and calibration models using bootstrapped Latin partitions. TrAC-Trends Anal Chem. 2006;25:1112–24. https://doi.org/10.1016/j.trac.2006.10.010.
65. Cruciani G, Baroni M, Clementi S, Costantino G, Riganelli D, Skagerberg B. Predictive ability of regression models. Part I: standard deviation of prediction errors (SDEP). J Chemom. 1992;6:335–46. https://doi.org/10.1002/cem.1180060604.
66. Wan C, Harrington P d B. Screening GC-MS data for carbamate pesticides with temperature-constrained–cascade correlation neural networks. Anal Chim Acta. 2000;408:1–12. https://doi.org/10.1016/S0003-2670(99)00865-X.
67. Harrington P d B. Multiple versus single set validation of multivariate models to avoid mistakes. Crit Rev Anal Chem. 2018;48:33–46. https://doi.org/10.1080/10408347.2017.1361314.
68. Harrington PB, Laurent C, Levinson DF, Levitt P, Markey SP. Bootstrap classification and point-based feature selection from age-staged mouse cerebellum tissues of matrix assisted laser desorption/ionization mass spectra using a fuzzy rule-building expert system. Anal Chim Acta. 2007;599:219–31. https://doi.org/10.1016/j.aca.2007.08.007.
69. de Boves HP. Support vector machine classification trees based on fuzzy entropy of classification. Anal Chim Acta. 2017;954:14–21. https://doi.org/10.1016/J.ACA.2016.11.072.
70. Aloglu AK, Harrington PB, Sahin S, Demir C. Prediction of total antioxidant activity of Prunella L. species by automatic partial least square regression applied to 2-way liquid chromatographic UV spectral images. Talanta. 2016;161:503–10. https://doi.org/10.1016/j.talanta.2016.09.014.
71. Rearden P, Harrington PB, Karnes JJ, Bunker CE. Fuzzy rule-building expert system classification of fuel using solid-phase microextraction two-way gas chromatography differential mobility spectrometric data. Anal Chem. 2007;79:1485–91. https://doi.org/10.1021/ac060527f.
72. Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AAM, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002;415:530–6. https://doi.org/10.1038/415530a.
73. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast Cancer. N Engl J Med. 2002;347:1999–2009. https://doi.org/10.1056/NEJMoa021967.
74. Guyon I, Li J, Mader T, Pletscher PA, Schneider G, Uhr M. Competitive baseline methods set new standards for the NIPS 2003 feature selection benchmark. Pattern Recogn Lett. 2007;28:1438–44. https://doi.org/10.1016/j.patrec.2007.02.014.
75. Bogdanov M, Matson WR, Wang L, Matson T, Saunders-Pullman R, Bressman SS, et al. Metabolomic profiling to develop blood biomarkers for Parkinson's disease. Brain. 2008;131:389–96. https://doi.org/10.1093/brain/awm304.
76. Abaffy T, Möller MG, Riemer DD, Milikowski C, DeFazio RA. Comparative analysis of volatile metabolomics signals from melanoma and benign skin: a pilot study. Metabolomics. 2013;9:998–1008. https://doi.org/10.1007/s11306-013-0523-z.
77. Bean HD, Jiménez-Díaz J, Zhu J, Hill JE. Breathprints of model murine bacterial lung infections are linked with immune response. Eur Respir J. 2015;45:181–90. https://doi.org/10.1183/09031936.00015814.
78. D'Amico A, Di Natale C, Paolesse R, Macagnano A, Martinelli E, Pennazza G, et al. Olfactory systems for medical applications. Sensors Actuators B Chem. 2008;130:458–65. https://doi.org/10.1016/j.snb.2007.09.044.
79. Franceschi P, Masuero D, Vrhovsek U, Mattivi F, Wehrens R. A benchmark spike-in data set for biomarker identification in metabolomics. J Chemom. 2012;26:16–24. https://doi.org/10.1002/cem.1420.
80. Schmekel B, Winquist F, Vikström A. Analysis of breath samples for lung cancer survival. Anal Chim Acta. 2014;840:82–6. https://doi.org/10.1016/j.aca.2014.05.034.