Talanta 116 (2013) 835–840

Contents lists available at ScienceDirect

Talanta

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

Evaluation of the effect of chance correlations on variable selection using Partial Least Squares-Discriminant Analysis

Julia Kuligowski [a,1], David Pérez-Guaita [b,1], Javier Escobar [a], Miguel de la Guardia [b], Máximo Vento [a,c], Alberto Ferrer [d], Guillermo Quintás [e,*]

[a] Neonatal Research Centre, Health Research Institute La Fé, 46009 Valencia, Spain
[b] Department of Analytical Chemistry, University of Valencia, 46100 Burjassot, Spain
[c] Division of Neonatology, University & Polytechnic Hospital La Fé, 46021 Valencia, Spain
[d] Department of Applied Statistics, Operations Research and Quality, Universidad Politécnica de Valencia, 46071 Valencia, Spain
[e] Leitat Technological Center, Bio In Vitro Division, 46021 Valencia, Spain

ARTICLE INFO

Article history:
Received 14 April 2013
Received in revised form
10 July 2013
Accepted 24 July 2013
Available online 2 August 2013

Keywords:
Metabolomics
Chance correlations
Variable selection
Partial Least Squares-Discriminant Analysis (PLSDA)

ABSTRACT

Variable subset selection is often mandatory in high throughput metabolomics and proteomics. However, depending on the variable to sample ratio there is a significant susceptibility of variable selection towards chance correlations. The evaluation of the predictive capabilities of PLSDA models estimated by cross-validation after feature selection provides overly optimistic results if the selection is performed on the entire set and no external validation set is available. In this work, a simulation of the statistical null hypothesis is proposed to test whether the discrimination capability of a PLSDA model after variable selection estimated by cross-validation is statistically higher than that attributed to the presence of chance correlations in the original data set. Statistical significance of PLSDA CV-figures of merit obtained after variable selection is expressed by means of *p*-values calculated by using a permutation test that included the variable selection step. The reliability of the approach is evaluated using two variable selection methods on experimental and simulated data sets with and without induced class differences. The proposed approach can be considered as a useful tool when no external validation set is available and provides a straightforward way to evaluate differences between variable selection methods.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

Nowadays, nuclear magnetic resonance (NMR) and the hyphenation of high resolution separation techniques (e.g. gas and liquid chromatography as well as capillary electrophoresis) with mass spectrometry (MS) play leading roles as high throughput analytical tools in comprehensive metabolomics and proteomics. Frequently, studies involve the discriminant analysis of samples under two distinct experimental conditions such as treated vs. untreated or diseased vs. control, for the identification of biomarkers or the calculation of predictive models. This is a challenging task, as the number of detected variables typically largely exceeds the number of samples, and variables are usually correlated. Moreover the unambiguous identification of metabolites or proteins can be highly difficult, and the concentration and response ranges involved normally cover several orders of magnitude. Besides, the majority of the detected variables are frequently irrelevant for the outcome prediction [1–3] and so the predictive precision and accuracy of discriminant models can be improved if uninformative variables are removed in advance [4,5]. Furthermore, feature selection also provides simplified models of easier interpretability in a subsequent statistical or biochemical data analysis.

Whereas a wide range of multivariate methods for supervised learning (i.e. pattern recognition) is available, each with its own strengths and weaknesses, the most commonly used multivariate classification technique is Partial Least Squares-Discriminant Analysis (PLSDA) [4]. PLSDA is a multivariate PLS method that extracts a set of latent variables (LVs) that explain the sources of variation in the $\boldsymbol{X}$-block correlated to a $\boldsymbol{y}$-vector that encodes the class membership [1]. One of the key features of PLSDA is its applicability in situations in which variables far outnumber samples, and correlation among variables exists [2,6]. In a PLSDA model, the relation between the predictors $\boldsymbol{X}$ $(N\times J)$ and the response $\boldsymbol{y}$ $(N\times 1)$ can be described as

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{b}^{T}+\boldsymbol{e} \qquad (1)$$

where $\boldsymbol{b}$ $(1\times J)$ is the vector of regression coefficients, $\boldsymbol{e}$ $(N\times 1)$ is the error vector (i.e. residuals) and N and J are the number of

[*] Corresponding author. Tel.: +34 937882300.
E-mail address: guillermo.r.quintas@uv.es (G. Quintás).
[1] Both authors contributed equally to this work.

0039-9140/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.talanta.2013.07.048

objects (e.g. MS or NMR spectra) and variables (e.g. *m*/*z* features or chemical shifts).

In metabolomic and proteomic studies, results should be subjected to thorough statistical and biological validation. Whereas the biological validation determines whether biomarkers are involved in processes related to the stated difference between classes, the statistical validation determines the performance of the biomarker and the probability of a chance result [7]. There are two statistical validation approaches, namely external and cross-validation. While external validation is considered the 'gold standard', cross-validation (CV) can be seen as a sub-optimal approximation to external validation [8] that, in spite of its limitations, still is very useful in case of a limited number of samples. Cross-validation is used for both the selection of the complexity of PLS models and to obtain an estimation of their predictive performance. During CV, a subset of objects from the data set is removed (i.e. validation set) and a PLS model is calculated using the remaining objects (i.e. training set). Then, the calculated model is used for the prediction of the $\boldsymbol{y}$ values of the validation set, and averaging over several splits yields the CV estimation of the model performance. CV methods are classified according to the procedure employed for the selection of the different subsets. In spite of being widely employed, CV increases the risk of model over-fitting and it also provides overoptimistic internal figures of merit in explorative and predictive PLS analysis [9,10]. Double cross-validation (2CV), also known as cross model validation, is an alternative CV strategy that circumvents these drawbacks providing external figures of merit [9–12]. In 2CV a subset of objects is set aside as a test set. The remaining set of objects are again split into training and validation sets, and they are subjected to a standard CV procedure for the selection of the number of latent variables [3]. Besides, non-parametric permutation tests based on random rearrangements of the elements of the $\boldsymbol{y}$ vector of a data set are useful for determining the significance of a statistic [9,13] and its use in combination with 2CV has been repeatedly proposed as a suitable approach to assess the statistical significance of PLSDA figures of merit [2,3,9,10,14].

As aforementioned, dimensionality reduction methods are often employed to increase PLS prediction accuracy. If variable selection is performed in advance on the entire data set, it gives overly optimistic CV results. This apparent improvement, however, partly originates from the susceptibility of variable selection towards chance correlations depending on both the variable to sample ratio and the correlation structure of data [15–18].

Addressing the aforementioned concerns, a straightforward strategy based on permutation testing and 2CV is proposed to grade the effect of chance correlations on PLSDA model performance during variable selection. For this purpose, the number of misclassified samples (NMC) and the discriminant Q^2 (dQ^2) [19] performance statistics calculated using real class labels were compared to a distribution of the same estimators obtained after class randomization before and after variable selection. Simulated data sets, an experimental MS data set and two variable selection procedures were used to demonstrate the potentials and drawbacks of the approach.

2. Material and methods

2.1. Software

Data analysis was run under Matlab 7.7.0 from Mathworks (Natick, USA, 2004) using in-house written MATLAB functions and the PLS Toolbox 6.2 from Eigenvector Research Inc. (Wenatchee, WA, USA). Bold capital letters represent matrices, bold italic lowercase characters represent vectors, and italic uppercase letters represent scalars. Both simulated and experimental data sets are assembled in matrices $\mathbf{X}$ ($N \times J$), where rows (N) and columns (J) correspond to samples and variables, respectively.

2.2. Data sets

Four 'null' data sets (N_{250} (60×250), N_{540}(60×540), N_{1000} (60×1000), and N_{2000}(60×2000)) were generated using the randn MATLAB function [20] and contained pseudorandom values drawn from standard normal distributions (i.e. mean zero and standard deviation one). The first 30 objects were classified as class A ($Y=1$) and the rest as class B ($Y=0$).

Then, a set of simulated data sets (**SIMUIN_5**, **SIMUIN_15** and **SIMUIN_25**) was calculated as described by Centner et al. [21]: **SIM** ($60 \times V$) was a simulated pure (noise free) data matrix generated with an exact dimensionality of 3, only containing informative variables. The **SIMUI** (60×540) data matrix resulted from the attachment of an uninformative variable matrix **UI** ($60 \times (540-V)$) to the **SIM** matrix. **UI** consisted of pseudorandom numbers drawn from a standard normal distribution. **SIMUIN** is the sum of the **SIMUI** matrix and a noise matrix **N** (60×540) containing pseudorandom numbers drawn from a normal distribution with mean zero and standard deviation 0.025. The **SIMUIN_5**, **SIMUIN_15** and **SIMUIN_25** data sets corresponded to $V=5$, 15 and 25, respectively. For each **SIM** matrix, a $\boldsymbol{y}$ (60×1) vector was calculated as $\boldsymbol{y}=3\boldsymbol{t_1}+2\boldsymbol{t_2}+1\boldsymbol{t_3}$, where $\boldsymbol{t_a}$ ($a=\{1, 2, 3\}$), is the vector of scores of the ath principal component. Class assignment of each simulated sample was carried out according to the sign of its calculated $\boldsymbol{y}$ value (see Fig. 1). A new $\boldsymbol{y}$ vector was generated where each class A sample was assigned a value of zero and each class B sample a value of one. The use of class labels (1/0) instead of the actual y values was selected to simulate real

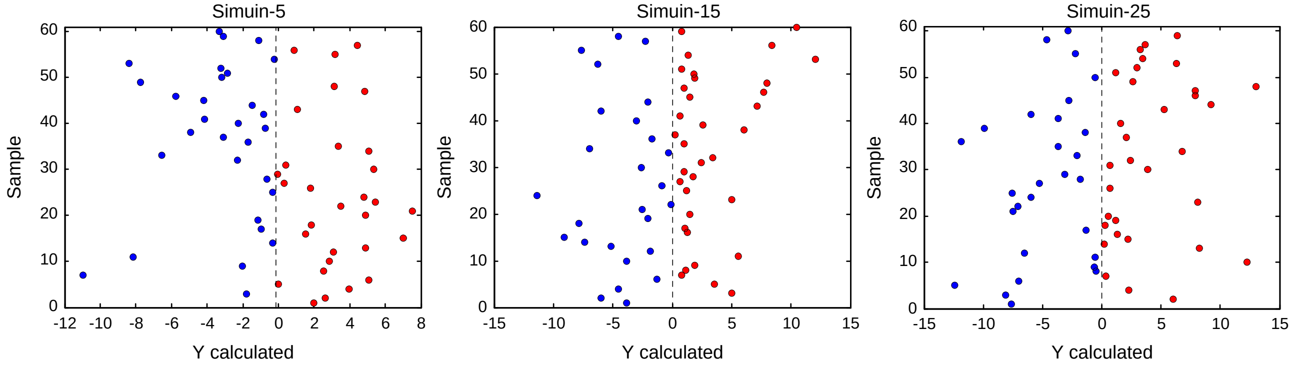

Fig. 1. Sample classification according to calculated y values of simulated data matrices. Note: blue circles: class A samples; red circles: class B samples; and dotted line: class threshold. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

situations where samples are typically clustered in two classes in spite of within-class differences among samples. Then, 20 randomly selected samples of each class were removed from each data set for being used as external test sets.

An experimental data set was used to test the applicability of the approach. The employed data set (**Gaucher** (40×590)) was obtained from the Biosystems Data Analysis Group website (www.bdagroup.nl) and contains Surface Enhanced Laser Desorption Ionization–Time of Flight–Mass Spectrometry (SELDI–TOF–MS) data of 40 serum samples from 20 Gauchy patients and 20 healthy controls. Each sample spectrum consists of 590 m/z variables between 1000 and 10,000. Y values of 1 and 0 were assigned to the spectra obtained from Gauchy and healthy patients, respectively. Background information on the **Gaucher** data set can be found in a previous work [22] and on the aforementioned website.

2.2.1. PLS modeling

Prior to PLS model calculation, autoscaling was employed to equal the relative importance of all variables. The $\boldsymbol{y}$ vector containing the class labels was mean centered. Scaling factors were calculated from the calibration subsets. No outlier detection was performed and all samples were used for variable selection and 2CV, employing a maximum of five PLS components selected from dQ^2 values calculated by CV.

2.3. Variable selection procedures

The following variable selection procedures were considered:

Approach 1. Variance of the $\boldsymbol{b}$ *regression vector* (b_{cv}*-PLSDA*):

This approach uses the set of PLSDA regression vectors obtained after M random K-fold cross-validations ($M=20$ and $K=4$ in this work). The number of LVs included in a PLSDA model was selected from dQ^2 values obtained after each K-fold CV. Then, the mean vector ($\bar{\boldsymbol{b}}$) and the standard deviation vector ($\boldsymbol{s_b}$) of the regression coefficient vectors ($\boldsymbol{b}$) were calculated. Variables were selected as informative according to ($|\bar{b}_j|-d\, s_{b,j}) > 0$ ($j=1,\ldots, J$). The value of d should be selected as a compromise between Type I and II errors. In this work $d=4$ was used.

Approach 2. Uninformative Variable Elimination (UVE-PLSDA):

The second approach involved UVE-PLSDA [21]. In this procedure, the original data matrix is augmented column-wise by a matrix containing normally distributed artificial random variables of very small magnitude (e.g. 10^9 times lower than the real variables [21]). Then, the standard deviation vector of the regression coefficients ($\boldsymbol{s_b}$) is obtained from the variation of the PLS regression coefficients by leave-one-out CV. The obtained values are used to calculate a reliability coefficient (c_j), which is an equivalent to the calculated t-value, for each original variable according to the following equation:

$$C_j = \frac{b_j}{\mathrm{std}(b_j)} \qquad (2)$$

Different criteria have been proposed [9] to establish the cut-off level for classification of real variables as informative using the reliability values of the artificial (uninformative) variables ($c_{artif,j}$). In this work, the UVE-R approach was employed where $|(c_{artif,j})|$ is ranked and a cut-off level corresponding to a defined α-quantile [21,23] is selected. Due to random generation of artificial variables, results found by UVE-PLSDA show certain variability as real variables with c_j values close to the cut-off level might be retained or not in the final model depending on minor differences in $c_{artif,j}$. Daszykowsky et al. [23] improved the performance of UVE-PLSDA by using a Monte Carlo approach in which UVE-PLSDA was repeated a number of times and in each run a randomly selected model set was used for model construction and feature selection.

The UVE-PLSDA procedure followed in this work can be described in four steps:

(i) *Matrix augmentation*: The original data matrix $\mathbf{X}$ was augmented column-wise by an artificial variable matrix $\mathbf{R}$ ($N \times 250$) with random values drawn from a standard distribution with mean zero and standard deviation 10^{-9} [21].

(ii) *Calculation of PLSDA submodels*: A series of N submodels for the augmented data matrix was calculated by leave-one-out CV. Accordingly, for each submodel, ($N-1$) samples were used to calculate an inner PLSDA model of complexities $a=\{1,\ldots, A\}$, which was subsequently used for prediction of the y_i value of the remaining sample. The procedure was repeated until all samples were predicted once as validation sample. The PLSDA model complexity was selected from dQ^2 values calculated using the predicted $\boldsymbol{y}$ values.

(iii) *Calculation of the reliability coefficient for each variable*: The reliability coefficient was calculated from the set of N regression vectors according to Eq. 2.

(iv) *Cut-off selection and UVE*: In this work, the α-quantile value of 99% was selected as cut-off value for a pre-classification of real variables as informative. To reduce the variability due to the use of random artificial variables during variable selection, the UVE-PLSDA process was repeated a total of 1000 times. By doing this, a frequency (γ) of pre-classification as informative was obtained for each variable. As informative variables are expected to be retained more frequently than uninformative variables, $\gamma=99\%$ was used as a second threshold value for variable discrimination.

The predictive performance of the PLSDA models after variable selection by both procedures was estimated by four-fold 2CV, as described elsewhere [9]. The random selection of 2CV training +validation and test sample subsets was repeated M times ($M=20$ in this work) to reduce the influence of the split on the results. Again, dQ^2 values calculated by leave-one-out CV within each training set were used to optimize the number of LVs of each inner model. Finally, average NMC and dQ^2 statistics were calculated from the obtained 2CV results.

2.4. Permutation test

In order to estimate the statistical significance of figures of merit obtained after variable selection, a permutation test was carried out to create a null distribution. Accordingly, the evaluation of the PLSDA performance using the selected variables was repeated 2000 times using randomly permuted class labels. p-Values for the figures of merit were calculated either empirically [3] or by tail approximation to a generalized pareto distribution (GPD) [13].

In the empirical approach, the p-value was computed as the fraction of permuted statistics that are at least as extreme as the test statistic obtained from the original data, as described elsewhere [3]. As the minimum Type-I risk after z iterations calculated this way is $1/z$, this empirical approximation becomes impractical when the calculation of each random statistic is computing intensive. In the second approach the (right) tail of the distribution of permutation values (i.e. x in Eq. (3)) using a maximum of 250 points is fitted to a generalized Pareto distribution with the following cumulative distribution function:

$$F(x) = 1-(1-kxa^{-1})^{1/k} \ , \quad \text{for } k \neq 0 \qquad (3)$$

This method is based on tail approximation and reduces the number of required permutations to accurately provide small *p*-values [13]. Detailed descriptions of the method, the data pretreatment and fitting procedure can be found in literature [13]. After estimation of both *k* and *a* parameters in Eq. (3), the Anderson–Darling statistic (A2) was calculated for the estimation of the goodness of fit of the data to a GPD [24]. If the test failed, the smallest exceedance was eliminated and the GPD fit was tested again. It has been demonstrated that this method provides accurate *p*-values with a reduced number of permutations as compared to the standard empirical approach. This advantage is of special importance when the number of permuted values exceeding the test statistic (*f*) is very low ($f \leq 10$, in this work) and the permutation approach is computing intensive. Nonetheless, in situations where the GPD fitting failed, the empirical *p*-value was used.

3. Results and discussion

3.1. Simulated data sets

First a study assessing the potential of the proposed approach to estimate the statistical significance of chance correlations on the improvement of PLSDA models after variable selection was performed using simulated data. Then the same approach was applied to the **Gaucher** data set.

3.1.1. Null data sets

Since predictors and responses were randomly generated in the null data sets, only non-statistically significant PLSDA models were expected before variable selection [2,3,9–11,15]. Accordingly, 2CV figures of merit (NMC and dQ^2) obtained before variable selection showed no class difference (e.g. NMC around 50% of the samples) for all four null data sets independently of their size (see Table 1). Nonetheless, the higher the variables-to-samples ratio, the higher the probability of finding a subset of variables with different distributions between classes because of sheer coincidence [7,15–19]. Consequently, after variable selection the number of variables selected as informative increased and figures of merit of the submodels improved with the numbers of variables in the original data set (see Table 1). For example, whereas the NMC for the null data set with 250 variables is reduced from 22 to 6 or 13, for the null data set with 2000 variables, the NMC can be artificially reduced from 24 to 0 after variable selection. This effect could be clearly seen using the b_{cv}-PLSDA approach where a correlation among the number of variables in the original data set, the number of retained variables and overoptimistic CV results was found. In spite of that, the permutation test showed the lack of statistical significance of the PLSDA submodels, expressed by the calculated *p*-values for both NMC and dQ^2 obtained using real class labels in comparison to those obtained from permutation testing, all giving *p*-values > 0.05.

This overoptimistic effect was further confirmed by results obtained for the simulated external test sets: whereas using variables selected by both b_{cv}-PLSDA and UVE-PLSDA approaches the number of misclassified samples employing cross-validation decreased rapidly (see Table 1), the number of misclassified samples in the external validation sets remained constant as shown in Table 2. For example, for the null data set (20 × 1000) the NMC in the external validation set before and after variable selection remains constant (equal to 11). Additionally, the number of selected variables for each null data set was close to the mean value of the number of retained variables using randomly permuted class labels. This can be appreciated from Figs. 2a and 3a for a **Null** (40 × 540) data set. Likewise, Figs. 2b and c, and 3b and c confirm that also NMC and dQ^2 values obtained for the same null data set are close to the mean value obtained for randomly permutated class labels, using both variable selection approaches.

In summary, results obtained from null data sets demonstrate that the evaluation of the statistical significance of figures of merit obtained after variable selection can be used to conclude whether there is a statistically significant difference between classes in the original data set. Nonetheless, this procedure is computing intensive and alternative approaches can also provide the same information faster with the same accuracy level [9].

3.1.2. SIMUIN data sets

Table 1 summarizes results obtained for the **SIMUIN5**, **SIMUIN15** and **SIMUIN25** data sets in which, while the variables-to-samples ratio was kept constant (540/40), the number of *a priori* informative variables increased from 5/540 up to 25/540. Results showed that the b_{cv}-PLSDA method retained percentages of *a priori* informative

Table 2
Number of misclassified (NMC) samples included in the test sets before and after variable selection.

	NMC before variable selection	NMC after variable selection	
		b_{cv}-PLSDA	UVE-PLSDA
Null (20 × 250)	12 (LV=2)	8 (LV=2)	12 (LV=1)
Null (20 × 540)	13 (LV=1)	10 (LV=1)	13 (LV=2)
Null (20 × 1000)	11 (LV=1)	11 (LV=1)	11 (LV=1)
Null (20 × 2000)	11 (LV=1)	7 (LV=1)	13 (LV=2)
SIMUIN_5 (20 × 540)	6 (LV=2)	4 (LV=2)	10 (LV=1)
SIMUIN_15 (20 × 540)	6 (LV=2)	3 (LV=1)	3 (LV=5)
SIMUIN_25 (20 × 540)	4 (LV=2)	2 (LV=2)	2 (LV=4)

Table 1
Figures of merit of PLSDA models established by 2CV and calculated for different data sets before and after variable selection. Standard deviations were obtained from four-fold 2CV results (see text for details).

Data set	Before variable selection		Variable selection method: b_{cv}-PLSDA			Variable selection method: UVE-PLSDA		
	NMC ± *s*	$dQ^2 \pm s$	Variables selected	NMC ± *s* (*p*-value)	$dQ^2 \pm s$ (*p*-value)	Variables selected	NMC ± *s* (*p*-value)	$dQ^2 \pm s$ (*p*-value)
Null (40 × 250)	22 ± 3	−0.22 ± 0.08	7	6 ± 1 (0.3)	0.54 ± 0.06 (0.3)	3	13 ± 1 (0.3)	0.06 ± 0.04 (0.3)
Null (40 × 540)	23 ± 3	−0.12 ± 0.07	23	1 ± 1 (0.1)	0.75 ± 0.02 (0.2)	9	13 ± 2 (0.3)	0.14 ± 0.04 (0.3)
Null (40 × 1000)	24 ± 3	−0.17 ± 0.06	34	1.6 ± 0.9 (0.9)	0.78 ± 0.03 (0.9)	5	16 ± 1 (0.7)	-0.02 ± 0.04 (0.7)
Null (40 × 2000)	24 ± 3	−0.14 ± 0.04	63	0.1 ± 0.4 (0.7)	0.87 ± 0.02 (0.8)	10	15 ± 1 (0.7)	0.09 ± 0.07 (0.8)
SIMUIN_5 (40 × 540)	15 ± 2	0.08 ± 0.05	26 (3/5)[a]	0.0 ± 0.2 (0.001)	0.87 ± 0.01 (0.001)	6 (0/5)[a]	14 ± 1 (0.7)*	0.16 ± 0.03 (0.5)*
SIMUIN_15 (40 × 540)	12 ± 2	0.20 ± 0.06	24 (8/15)[a]	0.4 ± 0.7 (0.01)*	0.82 ± 0.03 (0.02)*	9 (5/15)[a]	5.9 ± 0.7 (0.02)	0.58 ± 0.02 (0.004)*
SIMUIN_25 (40 × 540)	12 ± 2	0.31 ± 0.05	32 (15/25)[a]	1 ± 0.7 (0.05)*	0.82 ± 0.02 (0.01)*	11 (7/25)[a]	3.2 ± 0.4 (0.001)	0.67 ± 0.02 (0.001)
Gaucher (40 × 540)	6 ± 1	0.34 ± 0.07	52	0.1 ± 0.3 (0.001)	0.89 ± 0.03 (0.001)	47	9 ± 1 (0.007)	0.37 ± 0.02 (0.0035)

* *p*-value calculated using the GPD approach.

[a] Number of informative variables selected indicated in brackets.

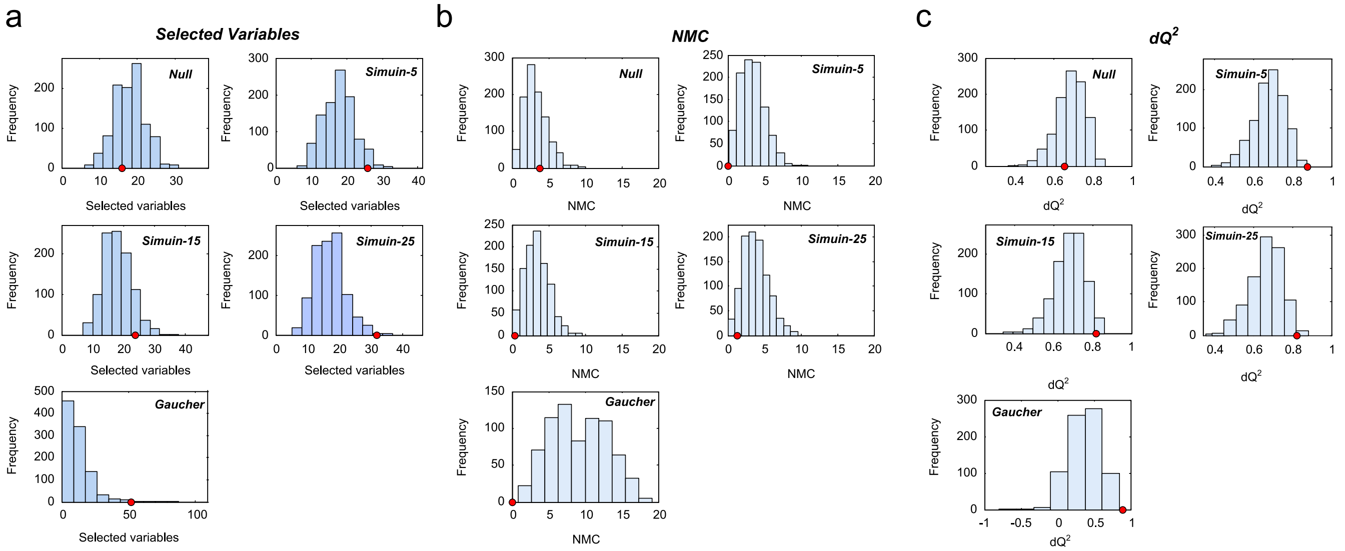

Fig. 2. Histograms of the number of selected variables (a), misclassified (NMC) samples (b) and discriminant Q^2 (c) in the simulated **Null** (40 × 540), **SIMUIN** and **Gaucher** data sets after variable selection using permuted class labels and the b_{cv}-PLSDA approach. Colored dots indicate values obtained using the original class labels.

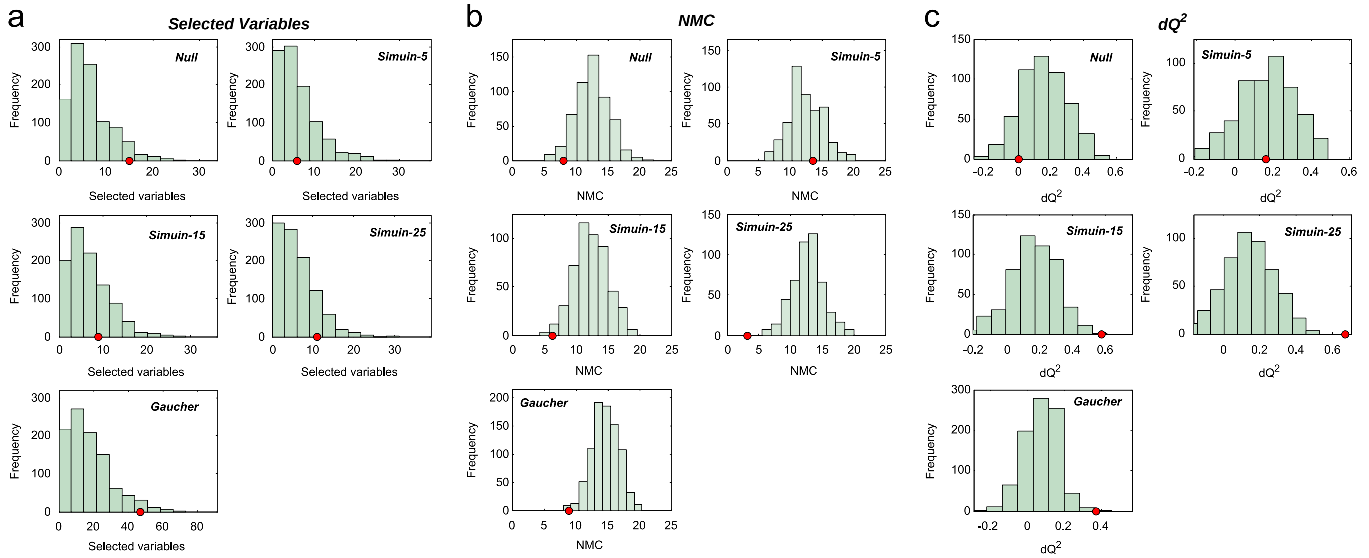

Fig. 3. Histograms of the number of selected variables (a), misclassified (NMC) samples (b) and discriminant Q^2 (c) in the simulated **Null** (40 × 540), **SIMUIN** and **Gaucher** data sets after variable selection using permuted class labels and the UVE-PLSDA approach. Colored dots indicate values obtained using the original class labels.

variables in the 53–60% range, and NMC as well as dQ^2 values was substantially improved after variable selection. Moreover, as depicted in Fig. 2 for the SIMUIN data sets, the number of variables retained was higher than those kept using randomly permuted class labels. Besides, statistically significant *p*-values were obtained for the figures of merit of the submodels thus indicating that the hypothesis that figures of merit using real and random class labels were equal could be rejected (α=0.05) and so, improvements were not exclusively due to existing chance correlations in the original data set. The suitability of both variable selection methods and the significance tests was also supported by lower NMC in the external validation sets after variable selection, as summarized in Table 2.

Likewise, results obtained from UVE-PLSDA for data sets **SIMUIN15** and **SIMUIN25** provided improved PLSDA figures of merit as shown in Tables 1 and 2 concerning the NMC and dQ^2 values obtained from 2CV as well as for the external validation set. Also results depicted in Fig. 3 are in good agreement with those obtained by b_{cv}-PLSDA showing the same trends in the number of selected variables, NMC and dQ^2 for SIMUIN data sets. Although the NMQ and dQ^2 values for **SIMUIN5** employing UVE-PLSDA had slightly improved after variable selection, indicating an improvement of submodel performance, *p*-values obtained for both, NMC and dQ^2 (see Table 1) indicated that the results obtained after variable selection were comparable to those due to chance correlations. This could also be confirmed by the NMC in the external validation set, increasing from 6 to 10 after variable selection. Indeed, whereas three out of five informative variables were selected using the b_{cv}-PLSDA approach, none of the those variables was retained by UVE-PLSDA. The observed differences between results found after b_{cv}-PLSDA and UVE-PLSDA selection in case of **SIMUIN5** were likely due to the effect of α and γ values on the set of retrieved UVE variables: whereas low values increase the number of both informative and uninformative variables retained, high thresholds may lead to a loss of useful information thus reducing the predictive capabilities of PLS models calculated after variable elimination.

3.1.3. Gaucher data set

The **Gaucher** (40 × 590) data set was obtained from a study focusing on the measurement of the protein profiles of serum of symptomatic Type I Gaucher patients (n=20) and controls (n=20) [22]. A total of 52 and 47 variables, 11 in common, were retained in the final models using b_{cv}-PLSDA and UVE-PLSDA, respectively. Results of this study using the two considered variable selection approaches provided p-values <0.05 for both dQ^2 and NMC as it can be seen in Table 1. In Figs. 2 and 3 it can be seen that the numbers of selected variables, NMC and dQ^2 values are different from the mean values obtained from permutation testing lying at the side of the random distributions as confirmed by the p-values shown in Table 1. It is interesting that all 10 variables identified in a previous work [22] as those with the largest contribution to the discrimination were included in both variable subsets.

When comparing figures of merit before and after UVE-PLSDA variable selection, results obtained where worse than it could have been expected. For example, the NMC after variable selection for the **SIMUIN_5** external test set was clearly worse than those found by using b_{cv}-PLSDA. The same effect was observed for the **Gaucher** data set where variable selection did not reduce the NMC.

Whereas the effect of chance correlations could not be eliminated (i.e. CV after variable selection provided overoptimistic figures of merit as it can be seen comparing the NMCs included in Tables 1 and 2), permutation testing provided a straightforward way to assess up to which extent the observed improvements in the predictive properties of PLSDA models after variable selection could be attributed to chance, and to compare different variable selection methods or conditions.

4. Conclusions

The elimination of variables irrelevant for classification is an important task that improves the predictive capabilities of multivariate models and facilitates their interpretation. Still, if the effect of chance correlations is unknown, variable selection must be performed in combination with an assessment of the obtained PLSDA models. Using simulated data sets as well as a real data set it could be shown that the inclusion of variable selection in the statistical validation process provides an estimation of its statistical significance, being useful when no external validation set is available. This procedure increases confidence in the variable selection process, which might be relevant for biological interpretation and development of further analysis methods (i.e. development of target methods) based on the obtained results. Furthermore, in spite of being computing intensive, this approach can also be useful to compare variable selection methods or conditions.

Acknowledgments

JE and JK acknowledge the “Sara Borrell” Grants (CD11/00154 and CD12/00667) from the Instituto Carlos III (Ministry of Economy and Competitiveness). DPG acknowledge the “V Segles” Grant provided by the University of Valencia to carry out this study. MV acknowledges the FISPI11/0313 Grant from the Instituto Carlos III (Ministry of Economy and Competitiveness). AF acknowledges the DPI2011-28112-C04-02 Grant from Spanish Ministry of Science and Innovation (MICINN). GQ acknowledges the financial support from the Spanish Ministry of Economy and Competitivity (SAF2012-39948).

References

[1] J.C. Lindon, J.K. Nicholson, E. Holmes, The Handbook of Metabonomics and Metabolomics, First ed., Elsevier, Amsterdam, 2007.
[2] K. Wongravee, G.R. Lloyd, J. Hall, M.E. Holmboe, M.L. Schaefer, R.R. Reed, J. Trevejo, R.G. Brereton, Metabolomics 5 (2009) 387–406.
[3] S. Smit, M.J. van Breemen, H.C.J. Hoefsloot, A.K. Smilde, J. Aerts, C.G. de Koster, Anal. Chim. Acta 592 (2007) 210–217.
[4] R. Brereton, Chemometric Pattern Recognition, First ed., Elsevier, Amsterdam, 2009.
[5] B. Simonetti, A. Lucadamo, M.R. González Rodríguez, Curr. Anal. Chem. 8 (2) (2012) 266–272.
[6] S. Wold, M. Sjostrom, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.
[7] S. Smit, Statistical Data Processing in Clinical Proteomics, Ph.D. dissertation, University of Amsterdam, Amsterdam, 2009.
[8] K.H. Esbensen, P. Geladi, J. Chemometrics 24 (2010) 168–187.
[9] J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven, F.A. van Dorsten, Metabolomics 4 (2008) 81–89.
[10] C.M. Rubingh, S. Bijlsma, E. Derks, I. Bobeldijk, E.R. Verheij, S. Kochhar, A.K. Smilde, Metabolomics 2 (2006) 53–61.
[11] P. Filzmoser, B. Liebmann, K. Varmuza, J. Chemometrics 23 (2009) 160–171.
[12] L. Gisdskeaug, E. Anderssen, B.K. Alsberg, Chemom. Intell. Lab. Syst. 93 (2008) 1–10.
[13] T.A. Knijnenburg, L.F.A. Wessels, M.J.T. Reinders, I. Shmulevich, Bioinformatics 25 (2009) I161–I168.
[14] G. Quintás, N. Portillo, J.C. García, J.V. Castell, A. Ferrer, A. Lahoz, Metabolomics 8 (1) (2012) 86–98.
[15] J.G. Topliss, R.J.J. Costello, J. Med. Chem. 15 (1972) 1066–1076.
[16] K. Baumann, N.J. Stiefl, J. Comput. Aided Mol. Des. 18 (2004) 549–562.
[17] K. Baumann, QSAR Comb. Sci. 24 (2005) 1033–1046.
[18] K. Baumann, Abstr. Paper Am. Chem. Soc. 227 (2004) U1026–U1027.
[19] J.A. Westerhuis, E.J.J. van Velzen, H.C.J. Hoefsloot, A.K. Smilde, Metabolomics 4 (2008) 293–296.
[20] Matlab R2011a, Mathworks, Natick, MA, 2011.
[21] V. Centner, D.L. Massart, O.E. de Noord, S. de Jong, B.M. Vandeginste, M.C. Sterna, Anal. Chem. 68 (1996) 3851–3858.
[22] S. Smit, M.J. van Breemen, H.C.J. Hoefsloot, A.K. Smilde, J.M.F.G. Aerts, C.G. de Koster, Anal Chim Acta 592 (2) (2007) 210–217.
[23] M. Daszykowski, I. Stanimirova, B. Walczak, F. Daeyaert, M.R. de Jonge, J. Heeres, L.M.H. Koymans, P.J. Lewi, H.M. Vinkers, P.A. Janssen, D.L. Massart, Talanta 68 (2005) 54–60.
[24] V. Choulakian, M.A. Stephens, Technometrics 43 (2001) 478–484.