The Monte Carlo validation framework for the discriminant partial least squares model extended with variable selection methods applied to authenticity studies of Viagra® based on chromatographic impurity profiles

B. Krakowska[1], D. Custers[2,3], E. Deconinck[2], M. Daszykowski[1,*]

[1] *Institute of Chemistry, The University of Silesia, 9 Szkolna Street, 40-006 Katowice, Poland*

[2] *Scientific Institute of Public Division of Food, Medicines and Consumer Safety, Section Medicinal Products Health (WIV-ISP), Rue Juliette Wytsmanstraat 14, Brussels, 1050, Belgium*

[3] *Research group NatuRA (Natural products and Food - Research and Analysis), Department of Pharmaceutical Sciences, University of Antwerp, Universiteitsplein 1, B-2610 Wilrijk, Belgium*

* Corresponding author: M. Daszykowski (e-mail: mdaszyk@us.edu.pl)
Tel.: +48 32 359 1568, Fax: +48 32 259 99 78

ABSTRACT

The aim of this work was to develop a general framework for the validation of discriminant models based on the Monte Carlo approach that is used in the context of authenticity studies based on chromatographic impurity profiles. The performance of the validation approach was applied to evaluate the usefulness of the diagnostic logic rule obtained from the partial least squares discriminant model (PLS-DA) that was built to discriminate authentic Viagra® samples from counterfeits (a two-class problem). The major advantage of the proposed validation framework stems from the possibility of obtaining distributions for different figures of merit that describe the PLS-DA model such as, e.g., sensitivity, specificity, correct classification rate and area under the curve in a function of model complexity. Therefore, one can quickly evaluate their uncertainty estimates. Moreover, the Monte Carlo model validation allows balanced sets of training samples to be designed, which is required at the stage of the construction of the PLS-DA and is recommended in order to obtain fair estimates that are based on an independent set of samples.

In this study, as an illustrative example, 46 authentic Viagra® samples and 97 counterfeit samples were analyzed and described by their impurity profiles that were determined using high performance liquid chromatography with the photodiode array detection and further discriminated using the PLS-DA approach. In addition, we demonstrated how to extend the Monte Carlo validation framework with four different variable selection schemes: the elimination of uninformative variables, the importance of a variable in projections, selectivity ratio and significance multivariate correlation. The best PLS-DA model was based on a subset of variables that were selected using the variable importance in projection approach. For an independent test set, average estimates with the corresponding standard deviation (based on 1,000 Monte Carlo runs) of the correct classification rate, sensitivity, specificity and area under the curve were equal to 96.42% ± 2.04, 98.69% ± 1.38, 94.16% ± 3.52 and 0.982 ± 0.017, respectively.

KEYWORDS

discriminant analysis; variable selection; counterfeit medicines; Viagra

1. INTRODUCTION

Models that are constructed using a large number of explanatory variables are prone to the overfitting issue. Namely, they tend to provide very optimistic predictions for the samples used to construct a model, but perform poorly for future samples, the so-called test set. This phenomena has been well-documented in the literature (see e.g. [1]) and is often observed when data consisting of instrumental signals are modeled. To deal with the overfitting issue two strategies are usually applied: (*i*) a careful selection of model complexity and (*ii*) the selection of relevant variables that support the construction of an adequate model. Both require appropriate validation.

In order to validate the performance of a model it is necessary to gain insight into its predictive abilities.[2] This can be done using different cross-validation/re-sampling methods (e.g. leave-one-out, leave-n-out cross-validation, bootstrap, jackknifing, Monte Carlo, etc.) and/or an independent set of samples. These model validation concepts are often referred to as internal and external validation. The cross-validation approaches are typically used to estimate the optimal number of components that should be used for the construction of a model. As claimed by Esbensen and Geladi[2], the only proper model validation is based on an independent test set. The test set is fully independent – it is never used at any stage of the construction of the model. It can be designed using subset selection approaches, for instance, uniform subset selection methods like Kennard and Stone algorithm and Duplex, clustering methods, random selection, D-optimality criterion, etc.[3] The general principles of multivariate model validation can be followed regardless of the intended purpose of a model – calibration, discrimination and/or classification.

In this study, we focused on the validation of the discriminant partial least squares models (PLS-DA) used in the field of drug control to discriminate authentic and counterfeit medicines based on their impurity profiles. The major motivation of our research was driven by the fact that in the literature that deals with the detection of counterfeit medicines, the role of model validation is often somewhat underestimated or even neglected. Assuming that authentic samples always have much lower levels of impurities compared to counterfeit samples PLS-DA seemed to be a straightforward choice that would allow the many correlated variables that are found in chromatographic and spectroscopic signals (fingerprints) to be dealt with. We introduced a general validation framework that is based on the Monte Carlo approach. In the course of the Monte Carlo procedure, distributions of selected figures of merit (e.g., sensitivity, selectivity, correct classification rate, etc.) are obtained from a discriminant model in the function of its complexity. Therefore, one can easily evaluate the performance of models that have different degrees of complexity based on the selected figures of merit and the uncertainties of their estimates. The proposed model validation procedure was especially designed to fulfill the PLS-DA assumption of balanced model sets. Moreover, we also assumed that model validation should be performed using balanced test sets in order to obtain error estimates for the groups of authentic and counterfeit samples. We also show that the validation scheme can easily be extended by the selection of the relevant variables that play a key role in differentiating authentic and counterfeit Viagra® samples (including the elimination of uninformative variables[4], the importance of a variable in the projections[5], selectivity ratio[6] and significance multivariate correlation[7]).

To illustrate the performance of the validation approach, we focused on the issue of the detection of authentic and counterfeit Viagra® samples based on their chromatographic fingerprints of impurities, which were determined using high performance liquid chromatography. A total of 143 samples of Viagra® medicines were analyzed, including 46 authentic and 97 counterfeit samples. The impurity fingerprints were further modeled using PLS-DA with the goal of developing a reliable diagnostic discriminant model.

2. THEORY

2.1 Preprocessing of chromatographic impurity fingerprints

Different preprocessing techniques are usually used to enhance the quality and interpretation of chromatographic impurity fingerprints. In addition to the peaks that originate from the analytes of interest, impurity profiles contain noise and baseline components that can influence the comparative analysis. In the course of a chromatographic analysis of complex mixtures, it is difficult to obtain chromatograms that have baseline separated peaks. Therefore, an over-expressed baseline component is often observed in chromatographic signals. When necessary, the baseline of the signal has to be corrected. A large number of baseline correction methods can be used to perform this task. One of these is the penalized asymmetric least squares method (PAsLS), which has found numerous applications. A detailed description of the algorithm is provided in reference [8].

Another preprocessing issue of instrumental signals is related to the possible shifts of the corresponding peaks that can be observed for a collection of signals. The presence of peak shifts in chromatographic fingerprints can be induced, for instance, by fluctuations in the instrumental conditions or changes in the chemical composition of an eluent and/or samples. Correlation optimized warping, COW, is one of the many alignment methods that are frequently used to diminish the negative effect of peak shifts [9]. The alignment of chromatographic fingerprints is achieved by the linear stretching and compression of signal sections in such a way that the overall correlation coefficient between the aligned signal and a target signal is maximized [9]. The target signal is considered to be a template for the alignment and usually reflects the highest correlation coefficient with the remaining signals [10]. Once this is done, it is considered to be representative.

In the COW method, each signal is divided into the same number of sections and all of the sections have the same number of sampling points. In order to match corresponding peaks, their shapes are evaluated. The alignment is controlled by two parameters – the number of sections, N, into which the signals are divided and the slack parameter, s, which influences the flexibility of the alignment. The alignment power of COW can be adjusted to a large extent by modifying the N and s parameters. The major advantage of using COW to adjust peak shifts stems from its ability to preserve the area and shape of a peak.

2.2 Exploratory analysis of impurity fingerprints

Principal component analysis, PCA, is an unsupervised technique that is used to visualize and compress multivariate data [11]. It is usually applied to explore the structure of multivariate data by means of low-dimensional projections that enable groups of samples, local changes of data density and/or objects with unique chemical characteristics compared with the majority of the data to be revealed [12]. In the framework of PCA, a data matrix is represented as the product of the score and loading matrices that contain the so-called principal components, PCs, in columns. It is important to stress that PCs are constructed in order to explain the largest part of the data variability.

Scores are linear combinations of the explanatory variables and are mutually orthogonal, whereas loadings are orthonormal, i.e. they are mutually orthogonal and have a unit length. A large absolute loading value of the original variable indicates the significant importance of that variable in the construction of a given principal component. Scores and loadings are used to visualize the data structure. Projections of selected pairs of scores and loadings provide information about any similarities among the samples and variables, respectively. The distances that are observed among samples (characterized by instrumental signals) in the dataspace described by the scores on selected principal components express their chemical similarity. Loading weight values, which are displayed on the loading plot, express the importance of the variables and the level of their mutual correlation.

2.3 Partial least squares discriminant analysis – model construction and validation

The partial least squares discriminant analysis, PLS–DA, is a variant of the classic partial least squares, which is built to distinguish samples from mutually exclusive groups in a linear manner [13]. For a two-class discriminant problem, a dependent variable, **y**, which drives the construction of the discriminant rule, defines the belongingness of a sample to a given group. For instance, for authentic vs. counterfeit samples, samples are usually labeled either as ‘-1’ and ‘+1’ or ‘0’and ‘1’.

In the course of model construction for a two-class problem, a logic rule is built using a few latent variables, which are constructed to maximize the description of data variance and at the same time maximize the covariance between a set of latent variables and a dependent variable. In the context of discrimination using the PLS-DA model, this objective is equivalent to minimizing the within-group scatter while maximizing the distance between the centers of groups [14].

The future performance of any discriminant or classification model is mostly affected by the representativeness of the samples that are used for its construction. These define its effective domain and influence model complexity (i.e. the number of latent PLS-DA variables). It is also important to emphasize that the construction of a PLS-DA model requires a balanced modeling set (i.e. one containing the same number of samples from each group). This issue was discussed in detail in [15]. It was proven that PLS-DA decision boundary is shifted towards a larger group, and thus it affects predictions of group labels. It becomes apparent that the samples that are used to construct a discriminant model, in fact, define its effective domain and thus should characterize the data variability expected during the model's maintenance. Usually, this is fulfilled by incorporating the most diverse samples selected from each group into the model set, which is based on a uniform subset selection approach (e.g., the Kennard and Stone algorithm or the Duplex algorithm [16],[3]).

The selection of the optimal number of PLS-DA latent variables that are necessary to obtain a model with satisfactory prediction properties can be guided by various cross-validation procedures[17]. The performance of a given model with respect to the recognition of the model and test set samples can be scored by several figures of merit, including the root mean square error for model set samples (RMSE) and the root mean square error for test samples (RMSEP). In general, these two figures of merit measure the level of the overall within-group scatter. However, other figures of merit are usually considered in discrimination problems. Among them the most popular are the correct discrimination/classification rate, CCR (the percentage or proportion of samples with a correctly predicted group label using a model) as well as sensitivity (SE) and specificity (SP). In order to obtain estimates of sensitivity and specificity, the number of true positive samples (TP), true negative samples (TN), false negative samples (FN) and false positive samples (FP) are evaluated. A sample from the group labeled '+1' is called a true positive sample when it is recognized as a sample from that group using a model; otherwise, it is a false negative. When a sample from the group labeled '-1' is recognized as a sample from that group using a model, it is called true negative, otherwise, it is a false positive. Sensitivity and specificity are defined as follows:

$$SE = TP / (TP + FN) \quad (1)$$

$$SP = TN / (TN + FP) \quad (2)$$

In order to illustrate the relation between the true positive and false positive rates and their influence on model parameters, the so-called ROC plots [18] were introduced. A model's performance is then expressed as the area under the convex curve, AUC. The larger it is the better the predictive properties of a given model, i.e. better discrimination power.

To test the reliability of a discrimination or classification model, it is possible to estimate the uncertainty associated with the estimates of certain figures of merit as a function of model complexity. This can be done in the course of the Monte Carlo [17] or the bootstrap procedure [19]. The Monte Carlo approach, MC, assumes the construction of many subsets by drawing samples in a random manner from the available groups. In this way, sources of variability can be simulated. At each step of the MC procedure, a subset of samples is selected and used to construct models with increasing complexity. Then, its figures of merit are calculated for each model using the remaining samples. Since the MC procedure is repeated many times, the distribution of a selected figure of merit can be constructed, thereby providing the possibility to obtain its uncertainty estimates. In the literature, different approaches for model optimization and validation are described, including the MC cross-validation, cross model validation, etc. [17],[20].

In this study, we adopted the idea of the MC validation specifically for a two-class discriminant problem that is solved using PLS-DA with an assumption of the balanced representation of two groups of samples (and also extended with the selection of relevant variables). A general scheme of the proposed MC validation procedure is presented in Fig. 1.

Fig. 1

At the first step, a balanced set of samples is selected at random from the available set of samples (the so-called initial model set). The remaining samples form an unbalanced set that is put aside as external test set. In the course of the MC procedure, the assumed number of samples is drawn randomly (without any replacement) from the initial model set and the pool of external test set samples. PLS-DA models with increasing complexity, 1, 2, …, *f*, are constructed for a given model subset. Each model is characterized by the selected figures of merit that are obtained for the balanced model set, internal test set and external test set. After j MC runs, the distribution for a given figure of merit and a fixed model complexity is obtained. The final results are reported as the average value, which is extended with the corresponding standard deviation as a function of model complexity. It should be emphasized that the external set of samples is independent with respect to all phases of modeling and variable selection, i.e. these samples are never used at the stage of model optimization or during variable selection – they only serve to test model performance.

Such a validation procedure allows the direct estimation of model complexity and straightforward validation using a completely external test set of samples. In order to enable a fair comparison and honest error estimates for two groups of samples, all of the subsets of the samples that are drawn in the course of the Monte Carlo approach are balanced. At the same time, model performance is revealed, which includes information about the uncertainty that is associated with the estimated figures of merit.

2.4 Selection of relevant variables for PLS-DA

Over the last few years, different variable selection approaches have been proposed in order to limit the risk of model overfitting and to enhance interpretation. Some of them, e.g., variable importance in projection (VIP) [5] and significance multivariate correlation (SMC) [7], are specifically designed to support variable selection for partial least squares regression and PLS-DA, whereas uninformative variable elimination and the selectivity ratio are not only limited to PLS [4],[21],[22].

2.4.1. Uninformative variable elimination-partial least squares discriminant analysis

Uninformative variable elimination partial least squares discriminant analysis, UVE-PLS-DA, is designed to eliminate variables that do not support the modeling of the dependent variable **y**. In practice, these variables contain information content that is comparable to the random variables [4]. To distinguish between informative and uninformative variables, the stabilities of the regression coefficients for the original variables and random variables (obtained in the course of the jack-knifing procedure) are compared. Therefore, the UVE-PLS-DA model is built for augmented data that contains the experimental data matrix $\mathbf{X}$ of dimensions $m \times n$ (m samples and n variables) and matrix $\mathbf{N}$ with random variables that are normally distributed with low magnitudes ($m \times n^*$). Artificial variables do not influence the construction of the UVE-PLS-DA model.

The stability of a regression coefficient is defined as the ratio between the mean value of the regression coefficients for a given variable obtained from the jack-knifing approach and their standard deviation. It is intuitive that informative variables must have absolute stabilities that are larger than the absolute values of the stabilities observed for the random variables. Therefore, the threshold value can be selected as, for instance, the maximal value of the absolute stabilities that are found for the random variables or a certain percentile value, e.g., 0.99.

Uninformative variables are then discarded and the final PLS-DA model is built based on the remaining variables.

2.4.2 Variable importance in projection

Variable importance in projection, VIP, is a variable selection method that helps to filter out variables in PLS-DA that are irrelevant for a given discriminant or calibration problem [23]. Importance of variables is scored by the VIP value, which is determined as described in reference [24]. The influence of variables is classified by means of the VIP score as follows: VIP > 1.0 (highly influential), 0.8 < VIP < 1.0 (moderately influential) and VIP < 0.8 (less influential). It is recommended that the VIP criterion be used in a recursive manner, i.e. to perform the elimination of variables until no further model improvement is observed.

2.4.3 Selectivity ratio

Another filter variable selection method that can be used to determine the importance of variables in PLS is the so-called selectivity ratio, SR. The selectivity ratio is calculated for each variable as the ratio between the variance explained by the PLS model and the variance of the model residuals [21],[6]. A high SR value means that a variable has a strong ability to discriminate the analyzed groups of samples. The threshold above which the variables are considered as important is arbitrarily chosen by the user.

2.4.4 Significance multivariate correlation method

The significance multivariate correlation method, SMC, assists in assessing the significance of variables in PLS regression or PLS-based discriminant analysis [7]. SMC belongs to the category of filter variable selection methods. Its main aim is to estimate the sources of the relevant variability for each variable based on model regression coefficients (i.e. explained variance and residual variance). The importance of variables is described by the SMC parameter, which is calculated for each variable taking into account the predicted response and the regression coefficients that are obtained from a given model. Variables with relatively high SMC values are better correlated with the response variable **y** and thus they can be regarded as relevant for a given regression or discrimination problem. In order to identify the relevant variables, the threshold value for the SMC values is defined based on the F-test with an assumed significance level of α and 1 and m - 2 degrees of freedom. Similar to VIP, the SMC is designed to assist in the assessment of the relevance of a variable in the construction of models that uses the PLS approach.

2.4.5 Variable selection and the Monte Carlo procedure

All of the variable selection approaches discussed above can be used during the MC procedure to obtain information about the relevance of variables and the frequency of their selection based on multiple model sets. A set of retained variables is stored at each step of the MC procedure. The final set of variables is determined for the model that has the optimal complexity assuming a certain frequency of their selection, e.g., variables found in 95% of all of the MC runs (see Fig. 2). Then, the final model that contains only relevant variables is constructed and validated using the general procedure shown in Fig. 1.

Fig. 2

3. EXPERIMENTAL

A collection of 46 authentic and 97 counterfeit Viagra® samples were analyzed using a high performance liquid chromatography system (Waters 2695 Separations Module, Milford, USA) with a photo-diode array detector (Waters 2998 Photodiode Array Detector, Milford, USA). The following sample preparation was applied – 30 mg of sample was dissolved in

10 mL of ethanol/water (50/50 % V/V) for 15 minutes using an ultrasonic treatment. Afterwards, the samples were centrifuged at a speed of 2,000 rpm for 10 minutes. The supernatant was used for chromatographic analysis. All of the steps of sample preparation were performed at room temperature.

Five µl of the sample was injected into the HPLC system. The autosampler temperature was 15°C and the column temperature was set to 30°C. Separation was carried out using a C18 column (Alltima, 250 mm × 3 mm; 5 µm particle size; Grace, Columbia, USA) and a binary mobile phase that was composed of an ammonium formate buffer (0.020 M, pH = 3) and methanol. Their proportions were controlled via the following gradient program – for two minutes; the mobile phase consisted of 90% of the buffer and 10% methanol. Then, the proportion of buffer and methanol were linearly altered to 50% for the next five minutes and kept at this level for the next seven minutes. Afterwards, the proportion of buffer and methanol reached 10% and 90%, respectively, for six minutes and this mobile phase composition remained constant for five minutes. During the last five minutes of the separation, the initial mobile phase composition was reached and the total separation run was completed after 30 minutes. A constant mobile phase flow rate of 0.5 ml·min^{-1}was applied during the analysis. Spectra were measured between 210 and 400 nm with 1.9 nm step for each portion of the eluent.

All data were acquired using the Empower software version 3 (Waters, Milford, USA). The final set that was used for the construction of the diagnostic models consisted of chromatograms recorded at the optimal detection wavelength (254 nm). At this wavelength excipients such as like lactose, croscarmellose, etc. do not absorb, and therefore they are not detected.

3.1 Hardware and software description

Calculations were performed using HP ProBook 6560b personal computer with processor Intel(R) Core(TM) i5-2520M CPU 2.50 GHz and 16 GB RAM, Operating System: Microsoft Windows 7 Version 6.1 (Build 7601: Service Pack 1). All discussed in this manuscript algorithms were developed in-house in the MATLAB computing environment (MATLAB Version: 8.1.0.604 (R2013a). The MATLAB routine for validation of PLS-DA using the MC framework and related algorithms are available from the corresponding author upon request.

4. RESULTS AND DISCUSSION

4.1 Data preprocessing

Since all of the chromatographic fingerprints contained the same number of sampling points (13,620) and were registered within the same range of elution times, they did not require resampling. In order to eliminate any differences in the baseline of the chromatographic fingerprints, the PAsLS method was used. The choice of the λ parameter was optimized. For most of the chromatographic fingerprints, $\lambda = 10^5$ was found to be the optimal value and the second degree of differences was considered.

The next step of preprocessing consisted of the alignment of the signals due to the presence of peak shifts. The misalignment issue was easily detected when the position of the active substance peak was analyzed across the collection of chromatograms. At the step of signal preprocessing, the peak of the active substance served as the marker peak, which additionally helped to verify the alignment. The correlation optimized warping method was used to correct the position of corresponding peaks. The target signal for the alignment was selected as the one that resembled the best average correlation with all of the remaining signals [10]. In order to achieve a satisfactory alignment using COW (in terms of the improvement of the correlation coefficient that was evaluated before and after alignment), different values of the input parameters were tested. In most cases, the values of ca. 28 sections ($N = 500$) and $s = 3$ enabled a relatively high alignment flexibility.

Afterwards, alignment tuning in two problematic signal sections was carried out (between 10.10 min to 10.43 min and between 20.10 min to 20.44 min) using the optimal values of the slack parameter that were equal to 3 and 6, respectively. For the first elution time, window N was equal to 20 and for the second 50.

The initial values of the correlation coefficients, which were calculated between the chromatographic fingerprints and the target signals, were in the range of 0.0134 to 0.9988. The initial correlation coefficients did not exceed 0.8 for 48.25% of the chromatograms and they were found to be below 0.9 for 64.34% of the chromatograms. A substantial improvement of peak correspondence was observed after signal alignment. 95.10% of the chromatograms were described by the final correlation coefficient above 0.8 and 90.21% of the chromatograms had a final correlation coefficient above 0.9. In order to visualize the alignment effect, histograms of the correlation coefficients that described individual chromatographic fingerprints before and after alignment are shown in Fig. 3.

Fig. 3

Only impurity profiles were considered in this study and therefore after the preprocessing step the peak of the active substance was removed from all signals (i.e. the region between 20.22 min and 20.85 min). Afterward, the chromatographic impurity fingerprints of the authentic and counterfeit Viagra® samples were further analyzed using the PCA approach and discriminated using PLS-DA.

4.2 Exploratory analysis of chromatographic impurity fingerprints

Potential differences between the authentic Viagra® samples and their counterfeit variants were explored using the PCA score projections that visualized the similarities among their impurity profiles. On the score projections in Fig. 4, authentic and counterfeit samples are marked as '+' and '○', respectively. More than 87.86% of the total data variance is explained by the first three principal components. Projection of samples onto a space that was defined by the first two principal components (Fig. 4a) revealed six unique (outlying) samples with very different chemical characteristics in comparison to the remaining samples. These belonged to the group of counterfeit samples. Due to outlying character of these samples, they were excluded from data set to eliminate their negative influence on the construction of the discriminant model.

In general, the analysis of score plots led to the conclusion that the group of counterfeit samples was much more inhomogeneous and scattered in comparison with the authentic Viagra® samples. This observation is not surprising since the production of counterfeit medicines has little to do with good manufacturing practice and maintaining reasonable levels of production quality. A larger scatter of counterfeit samples, especially along the PC 2 axis, confirmed the hypothesis that illegal counterfeiting practice is indeed a source of the additional variability that is manifested at the chemical level and can be readily explained by the increasing number and/or levels of impurities (see Fig. 4b). In addition, score projections indicated a separation tendency between the authentic and counterfeit samples. Apparently, the largest differences that were observed between authentic and counterfeit samples were due to the presence of the impurities that eluted at ca. 23.365 min and 2.855 min (see Fig. 4c). A larger scatter of counterfeit samples along PC 2 was mostly caused by a higher content of impurities that eluted at ca. 8.8 min (see Fig. 4d).

Fig. 4

4.3. Discrimination of authentic Viagra® samples and their counterfeit variants

In order to build a logic rule that could effectively support the discrimination of the authentic Viagra® samples and their counterfeit variants based on their chromatographic impurity profiles, the PLS-DA method was used. The selection of a linear discriminant method appeared to be a straightforward choice for our pilot study.

Regardless of the type of model that is considered, it is meant to support the decision-making process over a longer period of time. Its maintenance has very much to do with the samples used for its construction. In authenticity studies of different medicines, it is impossible to create or design a model set of samples since they will never reflect the potential variability of counterfeit variants. Therefore, local markets are sampled with the hope that the material that is acquired will describe the expected variability of counterfeit samples as well as possible. Construction of a model requires a representative set of samples that contain representative sets of authentic and counterfeit samples. Generally, the variability of authentic samples will be relatively small. Following this reasoning, in most of the published studies only a few authentic samples have been considered. On the other hand, the group of counterfeit samples is often considerably larger and this may raise two issues. The imbalanced proportions of samples will lead to difficulties in testing the constructed models since the independent test set will contain significantly fewer authentic samples than counterfeit samples. Using imbalanced groups of samples may, depending on the discriminant approach that is applied, influence the construction of the discriminant hyperplane. In support vector machines or k-nearest neighbor techniques, the construction of a separating hyperplane effectively involves only the samples that are located at the borders or very close to the borders of the groups. However, in LDA or PLS-DA, the situation is very different. All of the samples are required in order to determine the optimal location of the discriminant hyperplane and thus their proportions and group variances play a fundamental role.

In this study, we used a balanced model, test and external test sets of samples. The following number of samples were considered in the Monte Carlo scheme that is presented in Fig. 1: $m_1 = 46$, $m_2 = 97$, $p_1 = 35$, $h = 10$ and $g = 8$. All of the samples were selected randomly without any replacement. Such a construction of different sets offers the possibility to (*i*) design them with respect to the number of samples and (*ii*) simulate the different sources of variability. The MC procedure was repeated 1,000 times. In a single step, models with different number of factors were used in order to obtain estimates of the selected figures of merit (the correct classification rate, sensitivity, specificity and AUC) for a given configuration of the internal model set, test set and external test set.

After the MC procedure, a distribution of the selected figures of merit is available for a given model complexity, and thus their uncertainties can be estimated. Estimates of the correct classification rate are presented in Fig. 5, as the average value of correct classification rates extended with their standard deviation (vertical bars) from 1,000 MC runs as a function of model complexity.

Fig. 5

An analysis of the CCR values that were obtained in the course of the Monte Carlo procedure for internal test set suggested that the optimal PLS-DA model should contain 5 PLS factors, thus leading to 89.37% ± 1.48 of the correct classification rate for model set, 90.60% ± 3.97 for internal test set, and 88.03% ± 2.64 for external test set. Additionally, the relatively high values of sensitivity, specificity and AUC that were obtained for the external test set confirmed that the discriminant problem that was studied could be solved with a simple linear PLS-DA model (see Figs. 4b, c and d). A detailed presentation of all of the figures of merit that were considered is provided in Table 1.

Table 1

4.4 Variable selection

As was confirmed by the relatively high values of the figures of merit that are presented in Table 1, the PLS-DA model can differentiate authentic Viagra® samples and their counterfeit variants with great success. On the other hand, the large number of variables that were used for the construction of the model compared to the number of available samples increased the risk of model overfitting and complicated the identification of regions where the impurities, which are relevant for discrimination, eluted. Therefore, variable selection is usually recommended in order to limit model over optimism [25].

Several variable selection methods that can be easily embedded into the PLS-DA framework were considered to eliminate uninformative chromatographic features. These included uninformative variable elimination, variable importance in projection, the selectivity ratio and significance multivariate correlation. The final PLS-DA models were constructed using subsets of the relevant variables. Their performance was evaluated using the MC procedure, which was repeated 1,000 times and was tested with the external test set.

In the UVE procedure, only variables with absolute stabilities above the 99% percentile of the absolute stabilities were retained at each Monte Carlo step. The final set of variables contained variables that were identified as relevant in all of the MC runs. For VIP, only the variables selected in all of the runs were considered and the VIP procedure was applied recursively three times. Using the selectivity ratio, the threshold value was set to 0.9 and the final set of variables contained the variables selected in 95% of all MC runs (higher threshold values or higher selection frequencies resulted in an empty set of the selected variables). In SMC, the final set of variables contained variables characterized by 100% of selection frequency.

Analysis of the results that are presented in Table 1 allowed it to be concluded that, in general, the variable selection procedures that were applied decreased the complexity of all of the PLS-DA models. For a subset of variables selected using either SR or SMC, the final models contained one factor less compared to the initial model. The subsets of the variables selected with UVE and VIP resulted in a larger reduction of model complexity – from five to two factors. It is important to emphasize that the reduction of complexity for the studied data to have no negative effect on the prediction properties of the models (see Table 1). Of all of the variable selection methods that were applied to the data, SR had the most restrictive performance in terms of the number of discarded variables. Only 21 of the 13,291 variables were detected as relevant to the differentiation of authentic and counterfeit drug samples. The largest set of variables, which contained 3,641 variables, was retained using SMC. All of the constructed models, with and without a variable selection step, had average correct classification rates above 88% for the external test set with an uncertainty estimate below

2.86%. The best PLS-DA model, in terms of CCR, SE, SP and AUC estimates for the external test set, was constructed for the subset of variables selected using VIP (see Table 1). In general, one can conclude that regardless of the PLS-DA model with the VIP-based variable selection, all of the models had a tendency to describe the counterfeit samples better. This is supported by the larger specificity values compared with the corresponding selectivity values. On the other hand, the PLS-DA model with the VIP-based variable selection offered best performance for the authentic Viagra® samples (SE = 98.69% ± 1.38).

4. CONCLUSIONS

In this paper a general framework for the validation of PLS-DA models, which were built to discriminate authentic and counterfeit samples, was proposed. It takes into account balanced data representations based on the MC approach. Such an approach enables a simulation of variability due to the random selection of samples and at the same time the observation of model performance up to a maximal considered complexity. The major advantage stems from the possibility of obtaining distributions of figures of merit as a function of model complexity. It is also possible to easily extend the proposed framework using a variable selection procedure, e.g., VIP, SR, UVE or SMC. In general, such a strategy assists in reducing the over optimism of the PLS-DA models that are constructed and enhances their interpretation. The selected chromatographic variables (regions along the elution time axis) can help in the further identification of potential chemical markers of the counterfeiting processes that are studied using a complementary analytical technique. This strategy can also be used in other studies related to authenticity confirmation, which are designed to uncover differences between authentic and non-authentic samples at the chemical level, for instance to detect illegal fuel discoloration [26]. In general, all of the discussed PLS-DA models (with and without a variable selection scheme) offered a relatively high prediction performance. The best diagnostic model was based on PLS-DA constructed for a subset of variables selected using the variable importance in projection approach. The average estimates with corresponding standard deviations for the independent test set (based on 1,000 Monte Carlo runs) for correct classification rate, sensitivity, specificity and area under curve were equal to 96.42% ± 2.04, 98.69% ± 1.38, 94.16% ± 3.52 and 0.982 ± 0.017, respectively.

In general, the proposed validation workflow could also be used in many other discrimination and classification tasks, for instance, food adulteration or food authenticity studies (and not only) solved with other than PLS-DA discriminant models or class modelling techniques.

ACKNOWLEDGEMENTS

MD wishes to acknowledge the support of the National Science Centre, Poland (research grant no. 2014/13/B/ST4/05007)

FIGURE CAPTIONS

Fig. 1 A general validation scheme for partial least squares discriminant analysis based on the Monte Carlo procedure

Fig. 2 A general scheme of variable selection embedded into partial least squares discriminant analysis model construction based on the Monte Carlo procedure for the evaluation of variable selection frequency (fq)

Fig. 3 Histograms of the correlation coefficients that were calculated between a target signal and all of the chromatographic fingerprints: a) before and b) after alignment using the correlation optimized method

Fig. 4 Score plots of the first two principal components of impurity chromatographic fingerprints: a) 46 authentic '+' and 97 counterfeit '○' Viagra® samples, b) enlarged region of PC 1-PC 2 score plot, c) loadings on PC 1 with the three indicated elution regions at (1) 2.855, (2) 8.80 and (3) 23.365 min and d) loadings on PC 2 with four elution regions indicated at (1) 8.80, (2) 22.02, (3) 23.26 and (4) 23.37 min, where the most influential impurities elute

Fig. 5 a) correct classification rates (CCR), b) sensitivity (SE), c) specificity (SP) and d) area under curve (AUC) that were obtained as a function of the PLS-DA model complexity extended with uncertainty estimates expressed as standard deviation (vertical lines) obtained in the course of the Monte Carlo procedure (1,000 runs) for the internal model set (black line), internal test set (red line) and external test set (blue line)

REFERENCES

1 N. M. Faber and R. Rajkó, *Analytica Chimica Acta*, 2007, **595**, 98–106.

2 K. H. Esbensen and P. Geladi, *Journal of Chemometrics*, 2010, **24**, 168–187.

3 M. Daszykowski, B. Walczak and D. L. Massart, *Analytica Chimica Acta*, 2002, **468**, 91–103.

4 V. Centner, D.-L. Massart, O. E. De Noord, S. De Jong, B. M. Vandeginste and C. Sterna, *Analytical Chemistry*, 1996, **68**, 3851–3858.

5 O. M. Kvalheim, R. Arneberg, O. Bleie, T. Rajalahti, A. K. Smilde and J. A. Westerhuis, *J. Chemometrics*, 2014, **28**, 615–622.

6 O. M. Kvalheim, *J. Chemometrics*, 2010, **24**, 496–504.

7 T. N. Tran, N. L. Afanador, L. M. C. Buydens and L. Blanchet, *Chemometrics and Intelligent Laboratory Systems*, 2014, **138**, 153–160.

8 P. H. C. Eilers, *Anal. Chem.*, 2003, **75**, 3631–3636.

9 N.-P. V. Nielsen, J. M. Carstensen and J. Smedsgaard, *Journal of Chromatography A*, 1998, **805**, 17–35.

10 M. Daszykowski and B. Walczak, *Journal of Chromatography A*, 2007, **1176**, 1–11.

11 S. Wold, K. Esbensen and P. Geladi, *Chemometrics and Intelligent Laboratory Systems*, 1987, **2**, 37–52.

12 M. Daszykowski, B. Walczak and D. L. Massart, *Chemometrics and Intelligent Laboratory Systems*, 2003, **65**, 97–112.

13 M. Barker and W. Rayens, *J. Chemometrics*, 2003, **17**, 166–173.

14 E. K. Kemsley, *Chemometrics and Intelligent Laboratory Systems*, 1996, **33**, 47–61.

15 R. G. Brereton and G. R. Lloyd, *J. Chemometrics*, 2014, **28**, 213–225.

16 R. W. Kennard and L. A. Stone, *Technometrics*, 1969, **11**, 137–148.

17 Q.-S. Xu, Y.-Z. Liang and Y.-P. Du, *J. Chemometrics*, 2004, **18**, 112–120.

18 T. Fawcett, *Pattern Recognition Letters*, 2006, **27**, 861–874.

19 R. Wehrens, H. Putter and L. M. C. Buydens, *Chemometrics and Intelligent Laboratory Systems*, 2000, **54**, 35–52.

20 E. Szymańska, E. Saccenti, A. K. Smilde and J. A. Westerhuis, *Metabolomics*, 2012, **8**, 3–16.

21 T. Rajalahti, R. Arneberg, A. C. Kroksveen, M. Berle, K.-M. Myhr and O. M. Kvalheim, *Anal. Chem.*, 2009, **81**, 2581–2590.

22 M. Farrés, S. Platikanov, S. Tsakovski and R. Tauler, *J. Chemometrics*, 2015, n/a–n/a.

23 R. Gosselin, D. Rodrigue and C. Duchesne, *Chemometrics and Intelligent Laboratory Systems*, 2010, **100**, 12–21.

24 S. Favilla, C. Durante, M. L. Vigni and M. Cocchi, *Chemometrics and Intelligent Laboratory Systems*, 2013, **129**, 76–86.

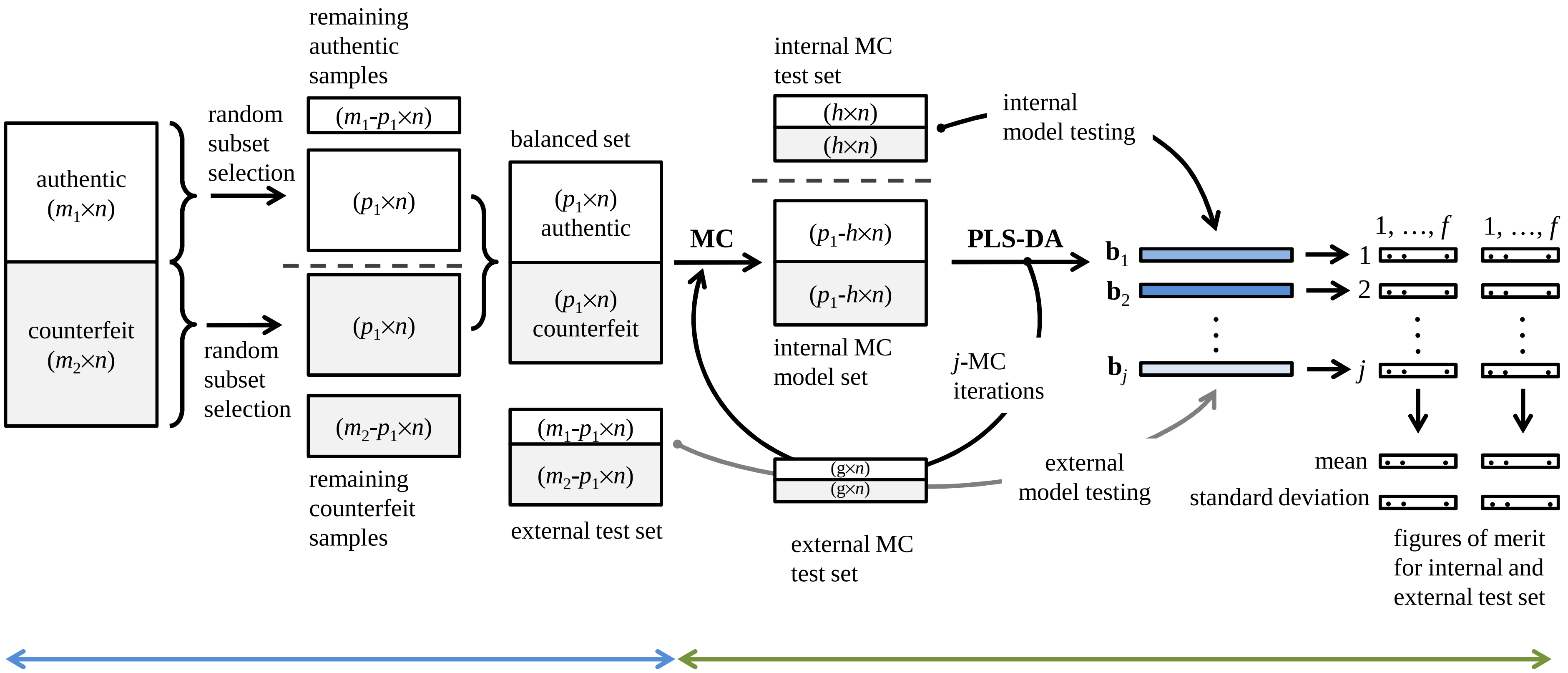
25 C. M. Andersen and R. Bro, *J. Chemometrics*, 2010, **24**, 728–737.

26 B. Krakowska, I. Stanimirova, J. Orzel, M. Daszykowski, I. Grabowski, G. Zaleszczyk and M. Sznajder, *Anal Bioanal Chem*, 2015, 1–12.

Table 1 Different figures of merit (CCR – correct classification rate, SE – sensitivity, SP – specificity and AUC – area under curve) that were obtained from the classic PLS-DA model and PLS-DA extended with four different variable selection schemes (UVE – uninformative variable elimination, VIP – variable importance in projection, SR – selectivity ratio and SMC – significance multivariate correlation) embedded into the Monte Carlo-based validation. Values of different figures of merit were obtained for model set samples, internal test set and independent test set samples randomly drawn 1,000 times in the course of Monte Carlo procedure. Each figure of merit is reported as the average value over 1,000 runs and accompanied by the estimate of uncertainty (standard deviation). Symbols *f* and *k* denote the number of latent PLS-DA factors and the number of considered (or selected using a variable selection technique) explanatory variables, respectively.

Model	*f*	*k*	Monte Carlo model set				Monte Carlo internal test set				Monte Carlo external test set			
			CCR [%]	SE [%]	SP [%]	AUC	CCR [%]	SE [%]	SP [%]	AUC	CCR [%]	SE [%]	SP [%]	AUC
PLS-DA	5	13,291	89.37 ± 1.48	82.82 ± 1.88	95.92 ± 3.25	0.999 ± 0.006	90.60 ± 3.97	82.44 ± 7.31	98.75 ± 4.06	0.984 ± 0.013	88.03 ± 2.64	82.48 ± 3.48	93.58 ± 5.03	0.972 ± 0.015
UVE	2	674	89.11 ± 1.44	82.47 ± 1.44	95.76 ± 2.62	0.960 ± 0.013	90.74 ± 3.70	81.64 ± 7.39	99.83 ± 0.56	0.891 ± 0.054	88.36 ± 2.19	82.45 ± 3.28	94.28 ± 3.40	0.940 ± 0.026
VIP	2	83	97.84 ± 1.39	99.10 ± 1.04	99.57 ± 2.26	0.999 ± 0.000	93.34 ± 3.32	100.00 ± 0.00	86.68 ± 6.65	0.956 ± 0.047	96.42 ± 2.04	98.69 ± 1.38	94.16 ± 3.52	0.982 ± 0.017
SR	4	21	91.32 ± 0.69	82.65 ± 1.37	100.00 ± 0.00	0.955 ± 0.008	90.71 ± 3.73	81.64 ± 7.39	99.79 ± 0.65	0.889 ± 0.046	89.72 ± 1.90	81.26 ± 3.40	98.19 ± 1.60	0.936 ± 0.022
SMC	4	3641	94.22 ± 1.95	91.47 ± 3.76	96.97 ± 2.79	0.986 ± 0.009	94.41 ± 3.11	90.30 ± 6.39	98.52 ± 2.58	0.999 ± 0.003	91.38 ± 2.86	88.71 ± 5.44	94.05 ± 4.35	0.962 ± 0.016

Fig. 1



Subset selection: random selection of a balanced set of samples and construction of external test set

Construction of the PLS-DA models with 1, 2, …, f factors in the course of the Monte Carlo procedure → figures of merit in a function of model complexity with their distributions

Fig. 2

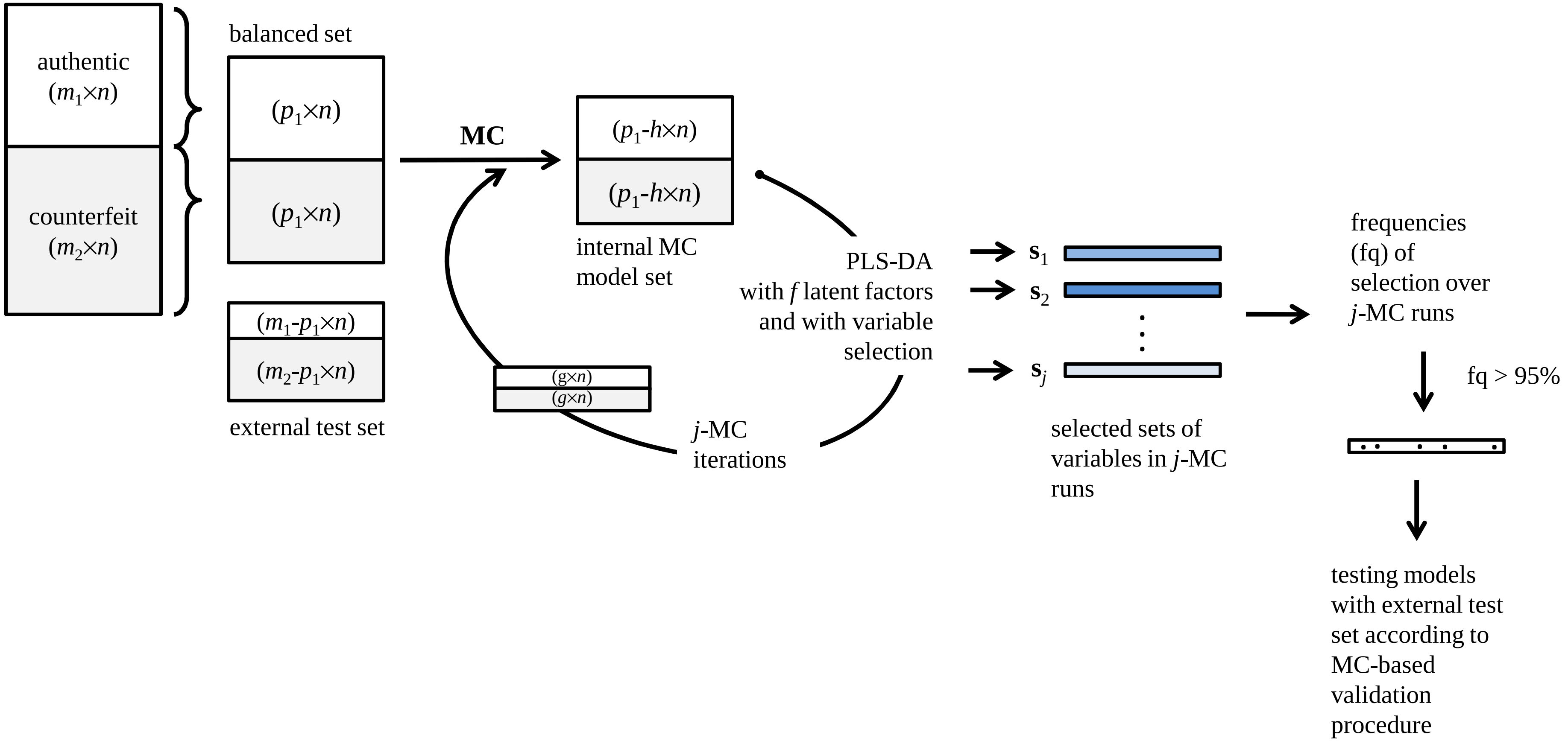

authentic
(m_1×n)
counterfeit
(m_2×n)
balanced set
(p_1×n)
(p_1×n)
(m_1-p_1×n)
(m_2-p_1×n)
external test set
MC
(p_1-h×n)
(p_1-h×n)
internal MC
model set
(g×n)
(g×n)
j-MC
iterations
PLS-DA
with *f* latent factors
and with variable
selection
$\mathbf{s}_1$
$\mathbf{s}_2$
$\mathbf{s}_j$
selected sets of
variables in *j*-MC
runs
frequencies
(fq) of
selection over
j-MC runs
fq > 95%
testing models
with external test
set according to
MC-based
validation
procedure

Figure 3

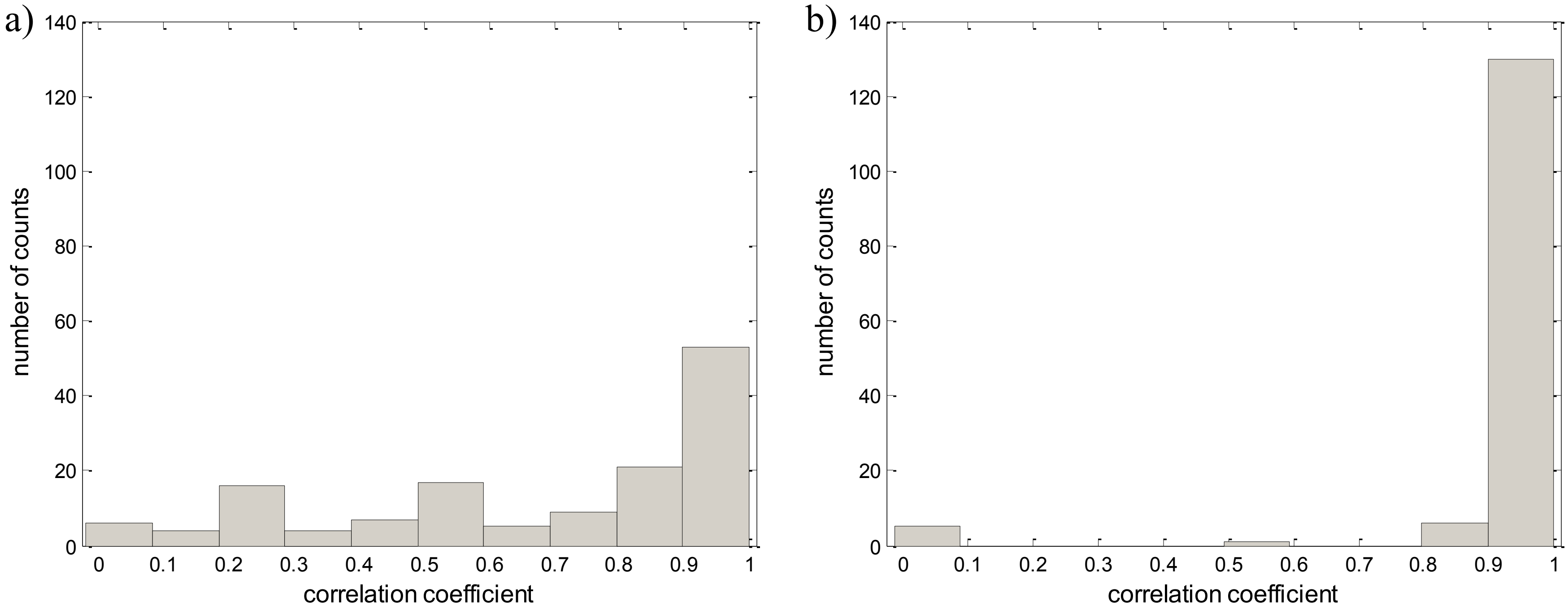

a)
b)
140
120
100
80
60
40
20
0
number of counts
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
correlation coefficient

Figure 4

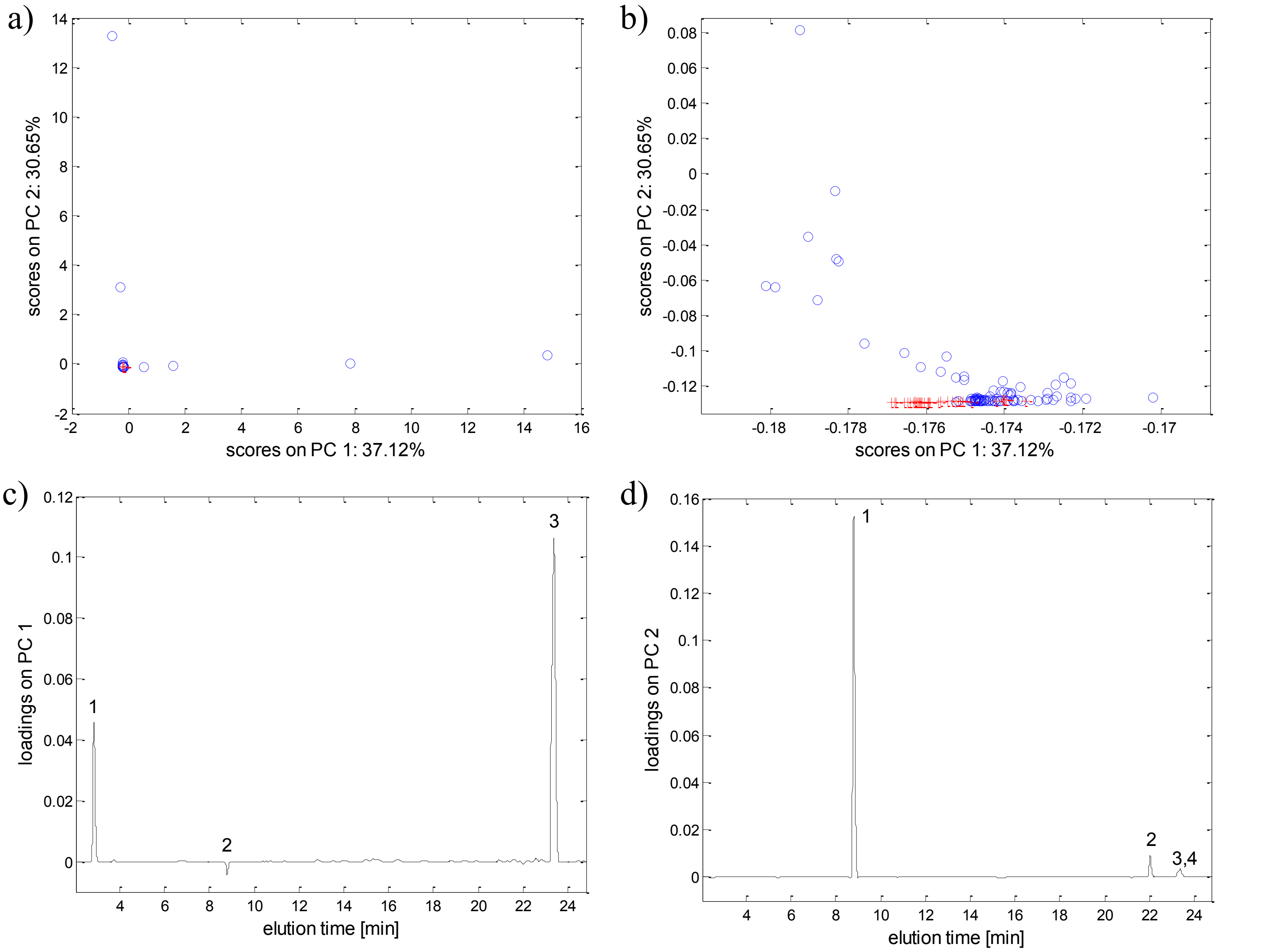
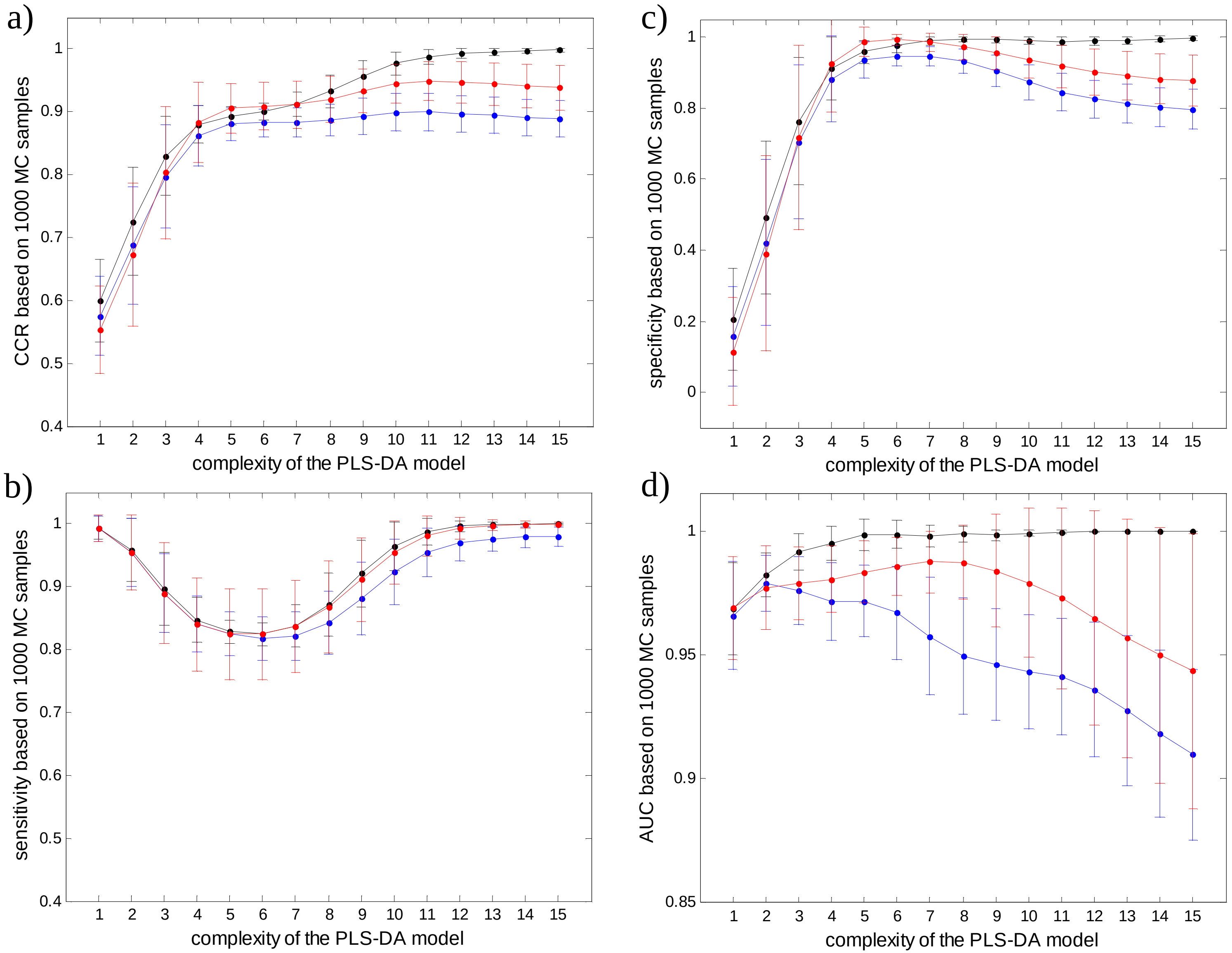

a)
scores on PC 2: 30.65%
14
12
10
8
6
4
2
0
-2
-2
0
2
4
6
8
10
12
14
16
scores on PC 1: 37.12%
b)
scores on PC 2: 30.65%
0.08
0.06
0.04
0.02
0
-0.02
-0.04
-0.06
-0.08
-0.1
-0.12
-0.18
-0.178
-0.176
-0.174
-0.172
-0.17
scores on PC 1: 37.12%
c)
loadings on PC 1
0.12
0.1
0.08
0.06
0.04
0.02
0
1
2
3
4
6
8
10
12
14
16
18
20
22
24
elution time [min]
d)
loadings on PC 2
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
1
2
3,4
4
6
8
10
12
14
16
18
20
22
24
elution time [min]

Fig. 5


a)
CCR based on 1000 MC samples
1
0.9
0.8
0.7
0.6
0.5
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
complexity of the PLS-DA model
b)
sensitivity based on 1000 MC samples
1
0.9
0.8
0.7
0.6
0.5
0.4
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
complexity of the PLS-DA model
c)
specificity based on 1000 MC samples
1
0.8
0.6
0.4
0.2
0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
complexity of the PLS-DA model
d)
AUC based on 1000 MC samples
1
0.95
0.9
0.85
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
complexity of the PLS-DA model