Anal. Chem. **2009,** *81,* 2581–2590

Discriminating Variable Test and Selectivity Ratio Plot: Quantitative Tools for Interpretation and Variable (Biomarker) Selection in Complex Spectral or Chromatographic Profiles

Tarja Rajalahti,[†,‡] Reidar Arneberg,[§] Ann C. Kroksveen,[∥] Magnus Berle,[∥] Kjell-Morten Myhr,[†,‡,⊥] and Olav M. Kvalheim*[,#]

Department of Clinical Medicine, University of Bergen, Bergen, Norway, Department of Neurology, Haukeland University Hospital, Bergen, Norway, Pattern Recognition Systems AS, Bergen, Norway, Institute of Medicine, University of Bergen, Bergen, Norway, The National Competence Centre for Multiple Sclerosis, Haukeland University Hospital, Bergen, Norway, and Department of Chemistry, University of Bergen, Bergen, Norway

The discriminating variable (DIVA) test and the selectivity ratio (SR) plot are developed as quantitative tools for revealing the variables in spectral or chromatographic profiles discriminating best between two groups of samples. The SR plot is visually similar to a spectrum or a chromatogram, but with the most intense regions corresponding to the most discriminating variables. Thus, the variables with highest SR represent the variables most important for interpretation of differences between groups. Regions with variables that are positively or negatively correlated to each other are displayed as corresponding negative and positive regions in the SR plot. The nonparametric DIVA test is designed for connecting SR to discriminatory ability of a variable quantified as probability for correct classification. A mean probability for a certain SR range is calculated as the mean correct classification rate (MCCR) for all variables in the same SR interval. The MCCR is thus similar to a mean sensitivity in each SR interval. In addition to the ranking of all variables according to their discriminatory ability provided by the SR plot, the DIVA test connects a probability measure to each SR interval. Thus, the DIVA test makes it possible to objectively define thresholds corresponding to mean probability levels in the SR plot and provides a quantitative means to select discriminating variables. In order to validate the approach, samples of untreated cerebrospinal fluid (CSF) and samples spiked with a multicomponent peptide standard were analyzed by matrix-assisted laser desorption ionization (MALDI) mass spectrometry. The differences in the multivariate spectral profiles of the two groups were revealed using partial least-squares discriminant analysis (PLS-DA) followed by target projection (TP). The most discriminating mass-to-charge (*m/z*) regions were revealed by calculating the ratio of explained to unexplained variance for each *m/z* number on the target-projected component and displaying this measure in SR plots with quantitative boundaries determined from the DIVA test. The results are compared to some established methods for variable selection.

Revealing the most discriminating variables in spectral or chromatographic fingerprints acquired for complex multicomponent samples represents a general analytical problem. The task is important, e.g., for finding biomarkers in profiles acquired for body fluids in proteomic and metabolomic/metabonomics studies.[1−4] A common approach to solve this kind of problem is to collect samples from the different groups and compare them using some kind of statistical tests or models. Linear discriminant analysis (LDA), which maximizes the between-group variance to within-group variance, was developed for this purpose by Fisher.[5] However, for problems where the number of variables is larger than the number of samples, LDA cannot be used without some kind of dimension reduction.[6] For such cases, principal component analysis (PCA)[7] is often used to look for discriminating patterns in multivariate data. PCA may provide decent results when the major variation in fingerprints represents between-group separation but is far from optimal when the major variation in the instrumental fingerprints is mostly shared for samples from different groups. When within-group variance dominates over between-group variance, methods that utilize a priori information

* To whom correspondence should be addressed. E-mail: Olav.Kvalheim@kj.uib.no. Phone: +47 55583366. Fax: +47 55589490.

[†] Department of Clinical Medicine, University of Bergen.

[‡] Department of Neurology, Haukeland University Hospital.

[§] Pattern Recognition Systems AS.

[∥] Institute of Medicine, University of Bergen.

[⊥] The National Competence Centre for Multiple Sclerosis, Haukeland University Hospital.

[#] Department of Chemistry, University of Bergen.

(1) Idborg-Björkman, H.; Edlund, P.-O.; Kvalheim, O. M.; Schuppe-Koistinen, I.; Jacobsson, S. P. *Anal. Chem.* **2003**, *75*, 4784–4792.

(2) Bijlsma, S.; Bobeldijk, L.; Verheij, E. R.; Ramaker, R.; Kochhar, S.; Macdonald, I. A.; van Omeen, B.; Smilde, A. K. *Anal. Chem.* **2005**, *78*, 567–574.

(3) Cloarec, O.; Dumas, M. E.; Craig, A.; Barton, R. H.; Trygg, J.; Hudson, J.; Blancher, C.; Gauguier, D.; Lindon, J. C.; Holmes, E.; Nicholson, J. *Anal. Chem.* **2005**, *77*, 1282–89.

(4) Jonsson, P.; Johansson, A. I.; Gullberg, J.; Trygg, J.; Grung, B.; Marklund, S.; Sjöström, M.; Antti, H.; Moritz, T. *Anal. Chem.* **2005**, *77*, 5635–5642.

(5) Fisher, R. A. *Ann. Eugen.* **1936**, *7*, 179–188.

(6) Barker, M.; Rayens, W. *J. Chemom.* **2003**, *17*, 166–173.

(7) Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52.

10.1021/ac802514y CCC: $40.75
Published on Web 02/19/2009

about the samples' group belongings show better performance. Partial least-squares discriminant analysis (PLS-DA)[8] is such a method. With two groups, a response variable is defined with zeros for the samples from one group and ones for the samples from the other group. Due to the use of a priori information about group belongings, PLS-DA performs better in discriminatory studies than PCA. This has been nicely demonstrated in a simulation study executed by Barker and Rayens[6] and was recently proven by Liu and Rayens.[9] However, similar to PCA, PLS-DA may lead to interpretational problems since the separation may require numerous PLS components, each one featuring the whole instrumental profile. For this reason many approaches to reduce the complexity of a PLS model by variable selection have been developed, using, e.g., the size of the covariance between the response and each instrumental variables (the so-called PLS weights),[10] size of regression coefficients,[11] the variable importance in projection (VIP) approach,[12] and selection of most predictive regions using a combination of genetic algorithms and PLS[13] or PCA.[14] Some of these approaches retain the "true" dimension of the model, whereas others reduce the model dimension by removing variance that is approximately orthogonal to the response.

Another route to solve the interpretational problem is to combine the PLS components into a single target-projected (TP) component.[15,16] The TP component represents the axis of maximum group discrimination for a PLS-DA model in the model space and thus the reduction of the PLS-DA model to a single predictive vector. This property provides simpler interpretation and is shared by the so-called orthogonal partial least-squares (OPLS) method.[17] Indeed, TP and OPLS only represent different algorithms to achieve the same predictive component.[18] The main objective of TP and OPLS is to overcome the interpretational problem posed by the orthogonal variation. Both TP and OPLS, however, are modeling the covariance between the instrumental variables and the response. Since variables large in absolute size usually also have large variances compared to variables with small absolute size, intense regions in spectra or chromatograms may dominate the model even if their variance is almost orthogonal to the response. Thus, since the variation in size between different spectral or chromatographic regions can be several orders of magnitude, the variables' intensity on the component of optimal group discrimination, in general, does not tell much about the variables' discriminatory ability. Wiklund et al.[19] tried to remedy the interpretational problem by introducing the *S*-plot, i.e., a plot of correlation versus covariance between the instrumental variables and the predicted response. However, the problem is only partially solved since the display becomes crowded when the number of spectral variables increases. Another possible solution to the interpretational problem is to scale the variables to equal variance before modeling. However, this approach may blow up noise since regions with little or no variance prior to scaling, i.e., low signal-to-noise (S/N) ratio, get the same variance as the most intense regions after scaling. An alternative solution to the interpretational problem was recently proposed by some of the present authors.[20] They calculated the ratio of explained to unexplained variance for each variable on the interpretative component and displayed these ratios similarly to a spectrum in the so-called selectivity ratio (SR) plot. In this plot, the most intense regions correspond to the variables with the best discriminatory ability. A problem that until now remained unsolved with the SR plot was to determine an objective limit to be able to quantitatively assess the statistical significance of a particular selection of discriminating variables. The present work addresses this problem and shows that the introduction of a nonparametric test to relate the selectivity ratio for each variable to mean correct classification rate (MCCR) provides a statistically founded threshold for variable selection. With the help of this discriminating variable (DIVA) test, the investigator can choose the probability level for his particular application to balance the risk of missing important discriminating variables against the possibility of including variables that result from chance correlations. This property is of special importance for applications where the variable-to-sample ratio is high. When plotting MCCR versus SR we obtain the DIVA plot; a new quantitative plot for aiding interpretation and variable (biomarker) selection.

In this work, we make another improvement to enhance the interpretative aspects of the SR plot: By multiplying the selectivity ratio with the sign of the corresponding regression coefficient, it is possible to quickly detect which variables are larger or smaller between groups. In metabonomics/metabolomics or proteomics applications it is then easy to comprehend which variables are up or down regulated.

THEORY

Latent-Variable Regression (LVR) of Instrumental Profiles. Let us assume that we have acquired instrumental profiles on samples from two groups. Each sample is characterized by a multicomponent profile of intensities at possibly tens of thousands spectral variables or chromatographic retention times. Our task is to decide which regions contain information with an ability to discriminate the two groups of samples. By introducing a response vector **y** of zeros and ones for the samples of the two groups, regression analysis can be used to solve this task. Partial least-squares discriminant analysis[8] or principal component regression (PCR)[21] can be performed in order to reveal variable regions that

(8) Sjöström, M.; Wold, S.; Söderström, B. In *Pattern Recognition in Practice II*; Gelsema, E. S., Kanal, L. N., Eds.; Elsevier: Amsterdam, The Netherlands, 1986; pp 461−740.
(9) Liu, Y.; Rayens, W. *Comp. Stat.* **2007**, *22*, 189–208.
(10) Höskuldsson, A. *Chemom. Intell. Lab. Syst.* **2001**, *55*, 23–38.
(11) Centner, V.; Massart, D. L.; de Noord, O. E.; de Jong, S.; Vandeginste, B. M.; Sterna, C. *Anal. Chem.* **1996**, *68*, 3851–3858.
(12) Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wold, S. *Multi- and Megavariate Data Analysis: Principles and Applications*; Umetrics: Umeå, Sweden.
(13) Nørgaard, L.; Saudland, A.; Wagner, J.; Nielsen, J. P.; Munck, L.; Engelsen, S. B. *Appl. Spectrosc.* **2000**, *54*, 413–419.
(14) Lavine, B. K.; Davidson, C. E.; Rayens, W. S. *Comb. Chem. High Throughput Screening* **2004**, *7*, 115–131.
(15) Kvalheim, O. M.; Karstang, T. V. *Chemom. Intell. Lab. Syst.* **1989**, *7*, 39–51.
(16) Kvalheim, O. M. *Chemom. Intell. Lab. Syst.* **1990**, *8*, 59–67.
(17) Trygg, J.; Wold, S. *J. Chemom.* **2002**, *16*, 119–128.
(18) Kvalhem, O. M.; Rajalahti, T.; Arneberg, R. *J. Chemom.* **2009**, *23*, 49–55.
(19) Wiklund, S.; Johansson, E.; Sjöström, L.; Mellerowicz, E. J.; Edlund, U.; Shockcor, J. P.; Gottfries, J.; Moritz, T.; Trygg, J. *Anal. Chem.* **2008**, *80*, 115–122.
(20) Rajalahti, T.; Arneberg, R.; Berven, F. S.; Myhr, K.-M.; Ulvik, R. J.; Kvalheim, O. M. *Chemom. Intell. Lab. Syst.* **2009**, *95*, 35–48.
(21) Jolliffe, I. T. *J. R. Stat. Soc., Ser. C* **1982**, *31*, 300–303.

discriminate between the two groups. In order to correct for noncompositional variation in the instrumental profiles, the profiles have to be pretreated before carrying out the regression analysis. The pretreated and centered matrix **X** with each row describing a sample and each column corresponding to intensities for one variable is decomposed into a product of two matrices and a residual matrix **E**. The two matrices are the orthogonal score matrix **T** and the loading matrix **P**.

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} = \mathbf{t}_1\mathbf{p}_1^{\mathrm{T}} + \mathbf{t}_2\mathbf{p}_2^{\mathrm{T}} + \ldots + \mathbf{t}_A\mathbf{p}_A^{\mathrm{T}} + \mathbf{E} \qquad (1)$$

Equation 1 also shows the alternative description of the latent-variable decomposition of **X** as a sum of products of score $\{\mathbf{t}_a\}$ and loading $\{\mathbf{p}_a\}$ vectors; $a = 1, 2, \ldots, A$. Superscript T implies transposition. Vectors are by default column vectors. Transposition transforms a column vector into a row vector. A is the number of extracted latent variables and is commonly determined by cross-validation.[22]

Target Projection. Whether PCR, PLS-DA, or other methods are used, the resulting decomposition usually consists of many components. This leads to difficulties for the interpretation of the model. Kvalheim and Karstang[15] developed a procedure called target rotation or target projection to simplify the interpretation of latent-variable regression models. Target projection produces a single predictive component by projecting the latent-variable decomposition onto the response variable. The target projection model can be written as

$$\mathbf{X} = \hat{\mathbf{X}}_{\mathrm{TP}} + \mathbf{E}_{\mathrm{TP}} = \mathbf{t}_{\mathrm{TP}}\mathbf{p}_{\mathrm{TP}}^{\mathrm{T}} + \mathbf{E}_{\mathrm{TP}} \qquad (2)$$

The subscript TP denotes the latent variable obtained by target projection. Scores and loadings for the TP model can be calculated from the decomposition given by eq 1.[15,16,18]

It has previously been shown[20] that the target-projected loadings are proportional to the product of the vector of regression coefficients $\mathbf{b}_{\mathrm{PLS}}$ and the covariance matrix $(\mathbf{X}^{\mathrm{T}}\mathbf{X})$. Thus, covarying variables on the discriminatory axis get enhanced loadings. This is an excellent property for the purpose of revealing variables strongly correlated to the response variable of group belongings: such variables will mutually strengthen each other. This is an important property since it means that discriminating variables small in absolute size will be enhanced due to their correlations with the other discriminating variables.

Variable Selection by Means of Discriminating Variable Test and Selectivity Ratio Plot. From eq 2, we can calculate explained $v_{\mathrm{expl},i}$ and residual $v_{\mathrm{res},i}$ variance for each spectral variable i in the TP model. From this we can define a selectivity ratio, SR, for each spectral variable i:

$$\mathrm{SR}_i = v_{\mathrm{expl},i}/v_{\mathrm{res},i} \qquad i = 1, 2, 3, \ldots \qquad (3)$$

The selectivity ratio can be displayed similarly to a spectrum.[20] The higher the value, the better the spectral variable discriminates between two groups of samples. Thus, the selectivity ratio can be used to quantitatively rank variables according to discriminatory ability. This is a valuable property for variable selection in general and maybe in particular for applications where the ratio of the number of variables to the number of samples is high. This situation is commonly encountered when we are searching for biomarkers in a complex profile of hundreds of chemical components. We are then facing the problem of being able to draw boundaries between probable, less probable, and improbable biomarker candidates. In other words: we need a tool to balance the possibility of including many false biomarker candidates against the possibility of missing important biomarkers.

A possible solution to define a boundary between variable regions with high discriminating ability and less interesting regions could be by comparing explained to residual variance in an F-test. In order to conclude that the variable has a high discriminatory ability, the explained variance on the TP component has to be significantly higher than the residual variance for a variable after removing the systematic variance explained by the TP component. This is the same as asking the question: for which variables have the introduction of the TP component explained enough variance to say that the variable has high discriminating ability? The answer of course depends on sample size N and chosen probability level α for the F-test. In order to reject the null hypothesis that explained and residual variance are the same, the calculated F value, F_{calc}, which is equal to SR_i from eq 3, has to exceed the critical value for the F distribution, F_{crit}.:

$$F_{\mathrm{calc}} = \mathrm{SR}_i > F_{\mathrm{crit}} = F_{(\alpha,N-2,N-3)} \qquad (4)$$

The number of degrees of freedom for the numerator (explained variance) in eq 3 is equal to the sample size N minus one degree of freedom due to the calculation of the variable's mean and one due to the introduction of the target component $(N - 2)$. For the denominator (residual variance) one extra degree of freedom is lost because we have to subtract the explained variance from the original variance of the variable. Thus, the remaining degrees of freedom for the denominator are $(N - 3)$.

Since F_{crit} converges toward one with increasing sample size N, an implicit assumption for this F-test is that selectivity ratios below one will not show good discriminatory ability. This is a rather strong assumption since it may well be possible that samples can be partially separated even if the predictive between-to-within group variance is significantly lower than one. Therefore, we propose to introduce a nonparametric test where the probability is derived directly from a measure of how well all variables within a certain SR interval separate two groups of samples. A good measure can be obtained by calculating the correct classification rate (CCR) for all variables. Completely random classification of the two groups on a variable corresponds to 50% CCR with equal number of samples in each group. On the other hand, if a variable separates the two groups completely, one group of samples is located on the low side of the values of that variable and the other group of samples is located on the upper side of the values. Such a variable is on top of the performance ladder with CCR of 100%. It is obvious that SR and CCR must be intimately correlated: increasing SR should provide increasing CCR since SR is a measure of the variables performance for separating groups. If we calculate CCR for all the variables, we can define selectivity ratio intervals and calculate an MCCR and its standard deviation for the variables in each SR interval and plot MCCR versus SR for the whole range of selectivity ratio intervals. This DIVA plot

(22) Bro, R.; Kjeldahl, K.; Smilde, A. K.; Kiers, H. A. L. *Anal. Bioanal. Chem.* **2008**, *390*, 1241–51.

provides an opportunity to objectively choose a threshold for discriminatory ability that balances the risk of missing important variables against the risk of selecting many variables resulting from chance correlations. (See the Results and Discussion for an example of a DIVA plot.)

From the nonparametric DIVA test we can obtain probability-based boundaries for the SR plot. This provides a quantitative display for assessing the discriminatory ability of all regions in a complex variable profile. Furthermore, we can take advantage of the fact that the sign of the regression coefficient for a variable shows if a variable increases or decreases between two groups of samples on the TP component. By multiplying the selectivity ratios with the sign of the corresponding regression coefficients, the SR plot quantitatively displays all important features for interpreting the target component and making an objective selection of discriminating variables (e.g., biomarkers).

DIVA Test and SR Plot for the Multiple Group Case. We have developed the theory for the DIVA test and SR plot for cases with two groups. Is it possible to generalize the approach to applications with multiple groups? First, we have to notice that there are different types of multiple group situations. For instance, in metabolomic or proteomic applications, we can have a group of healthy controls and groups of patients defined by different stages of the pathogenesis of the same disease, or we can have a group of healthy controls and several groups with different diseases. In the former case, we can define a single-response vector with values reflecting the development of the disease, e.g., zero for controls, one for first stage, two for second stage, and so on. The generalization of the DIVA test and the SR plot to handle this case is straightforward. An underlying assumption is that the different disease stages represent a continuous and gradual development that is reflected in the instrumental profiles. If this is not the case, a better solution may be to let the group of controls and a selected stage define the response and use the derived model to relate the pathogenesis to predicted responses for the other stages. In the latter case, where the signatures of different diseases have to be compared to healthy controls and the other disease, there is no mean to rank samples on a single response and pairwise comparison is the most optimal way of using the proposed approach. If prediction of group belonging for new samples is the aim, a possible continuation is to model each group independently on the selected features and fit new samples to each group to find the group providing best fit for each sample.[7]

Relation between ROC and the DIVA Plot. A receiver operating characteristics (ROC) or ROC curve is a bivariate plot of sensitivity versus (1 − specificity) in a binary classification. Sensitivity measures the fraction of actual positives which are correctly identified as such, whereas specificity measures the proportion of negatives which are correctly identified. Since the CCR is identical to sensitivity in a binary classification, the MCCR can be interpreted as a mean sensitivity for the variables within a certain SR interval. Thus, the DIVA plot connects classification performance to variables' ratio of between-to-within group variance in a quantitative manner and expands the ROC curve into the multifeature domain.

EXPERIMENTAL SECTION

Peptide Standard. Peptide calibration standard was purchased from Bruker Daltonics. The peptide calibration standard contained polypeptides with reference m/z values and names listed in Table 1. In the peptide standard, each polypeptide had a concentration of 4 pmol/μL.

Table 1. Composition of Peptide Standard Used for Spiking of CSF

m/z value	name
1047.20	angiotensin II
1297.51	angiotensin I
1348.66	substance P
1620.88	bombesin
2094.46	ACTH clip 1−17
2466.73	ACTH clip 18−39
3149.61	somatostatin 28

Samples. Cerebrospinal fluid (CSF) was drawn from patients undergoing spinal anesthesia for lower extremity orthopedic surgery. About 2−3 mL of CSF was drawn from each patient. The CSF was immediately centrifuged at 450g for 10 min to remove cells and thereafter stored at −80 °C. CSF from all individuals was randomly partitioned into five groups. One group labeled 0 pM was selected as reference. CSF from the other four groups was spiked with 50, 100, 200, or 400 pM peptide standard. Each sample was fractionated in duplicates through individual 30 kDa molecular weight cutoff (MWCO) filters. Triplicates of 5% of the resulting concentrated flow-through fractions were spotted onto the MALDI plate and analyzed. For further details on sample preparation prior to mass spectrometry (MS) profiling see Berven et al.[23]

Matrix-Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) Analysis. The low molecular weight (MW) fractions were analyzed using an AutoFlex (Bruker Daltonics) mass spectrometer in a positive linear mode. Data were acquired in the range of 740−9000 Da. The parameter settings for the mass spectral profiling of the low MW fraction were the following: laser frequency 20 Hz, ion source I 20 kV, pulsed ion extraction 250 ns with ion suppression up to 500 Da. No real-time smoothing was performed. The analyses were performed using the AutoXecute mode with the following setup: 20 initial uncollected shots at 35% laser power, followed by 100 shots that were collected. This was repeated at different positions until a total of 600 shots had been collected. The laser power was varying between 20% and 24% for the different experiments. The spectra were automatically collected if the signal-to-noise ratio was evaluated by the FlexControl software (version 2.0, Bruker Daltonics) in AutoXecute mode to be above 3, with a peak resolution of 200.

Data Sets. Each spectral profile acquired was described by intensities at 44 403 m/z numbers, starting at 740.04 Da and increasing in steps of 0.186 to 8999.84 Da. This provided a data set consisting of approximately 170 spectral profiles for the reference samples (0 pM) and approximately 50 spectra for each of the samples spiked with 50, 100, 200, or 400 pM.

(23) Berven, F. S.; Kroksveen, A. C.; Berle, M.; Rajalahti, T.; Fikka, K.; Arneberg, R.; Myhr, K.-M.; Vedeler, C.; Kvalheim, O. M.; Ulvik, R. J. *Proteomics—Clin. Appl.* **2007**, *1*, 699–711.

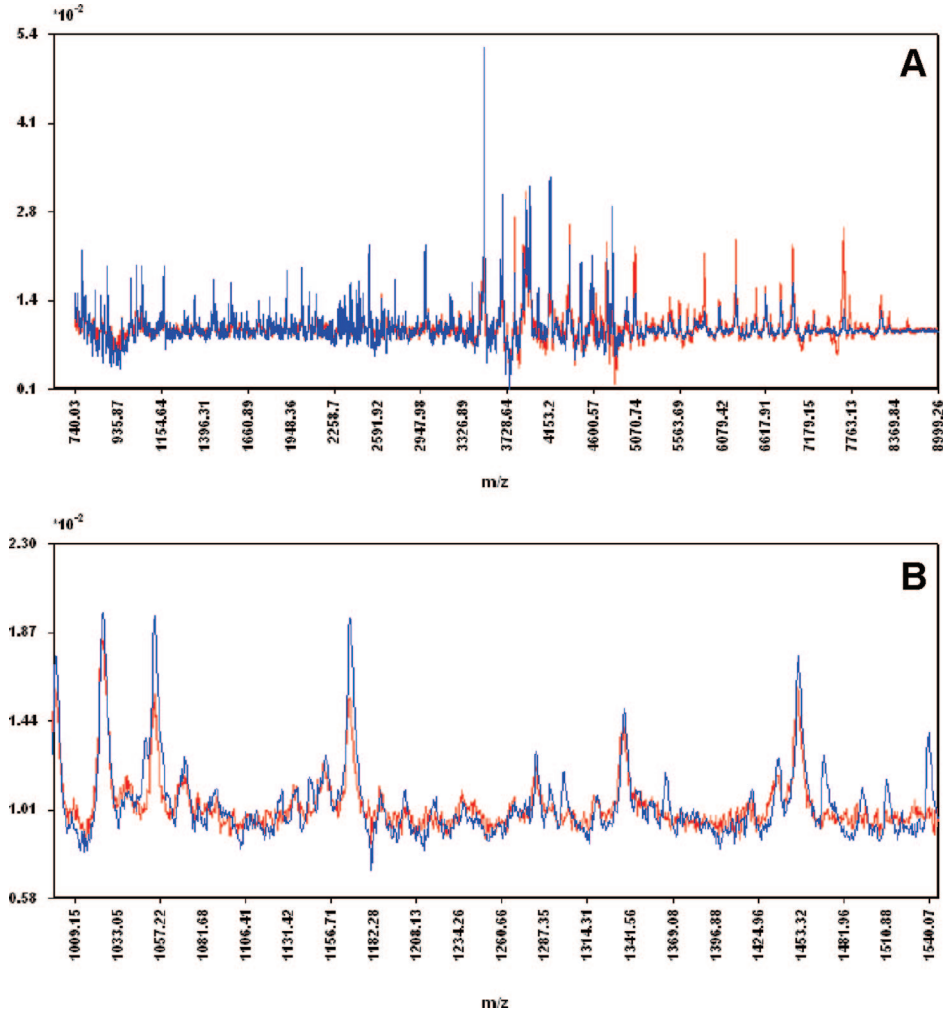

Figure 1. (A) Spectral profiles for the mass range of 740−9000 Da after pretreatment (8881 binned *m/z* numbers): reference CSF sample in red, CSF sample spiked with 100 pM peptide standard in blue. (B) Zoomed profiles.

Table 2. Modeling Results for the Different Sample Groups (0 pM vs 50−400 pM)

group (pM)	no. of spectra	A^a	$R^2(\mathbf{X}_{PLS\text{-}DA})^b$	$R^2(\mathbf{X}_{TP})^c$	$R^2(\mathbf{y})^d$
400	50	12	71.9	3.0	98.2
200	51	12	71.8	2.3	98.1
100	53	12	73.1	4.8	98.1
50	53	12	72.1	2.2	97.6

[a] A = number of components. [b] $R^2(\mathbf{X}_{PLS\text{-}DA})$ = explained variance in **X** for PLS-DA model. [c] $R^2(\mathbf{X}_{TP})$ = explained variance in **X** for TP model. [d] $R^2(\mathbf{y})$ = explained variance in *y* for PLS-DA and TP model.

Data Analysis/Pretreatment. The data were pretreated in accordance with the recommendation by Arneberg et al.[24] Baseline correction was performed using the FlexAnalysis software from Bruker Daltonics. This produced some regions with negative intensities, and therefore the profiles were independently shifted by the absolute value of the largest negative intensity in each profile prior to further pretreatment. Prior to alignment, binning was performed by adding the intensities of five consecutive *m/z* numbers. This reduced the number of variables from 44 403 to 8881 and, in addition, provided a smoothing of the spectra. Alignment was executed by using the algorithm of Wong and

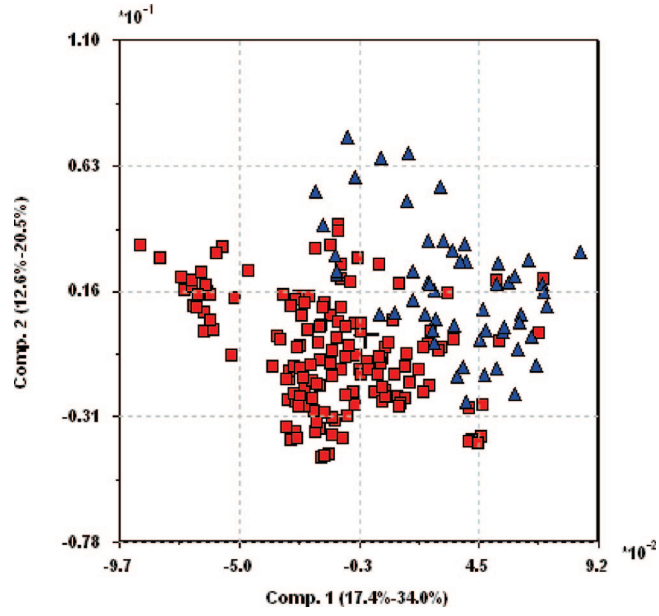

Figure 2. PLS-DA score plot for reference samples (red) and samples spiked with 100 pM peptide standard (blue).

(24) Arneberg, R.; Rajalahti, T.; Flikka, K.; Berven, F. S.; Kroksveen, A. C.; Berle, M.; Myhr, K.-M.; Vedeler, C.; Ulvik, R. J.; Kvalheim, O. M. *Anal. Chem.* **2007**, *79*, 7014–7026.

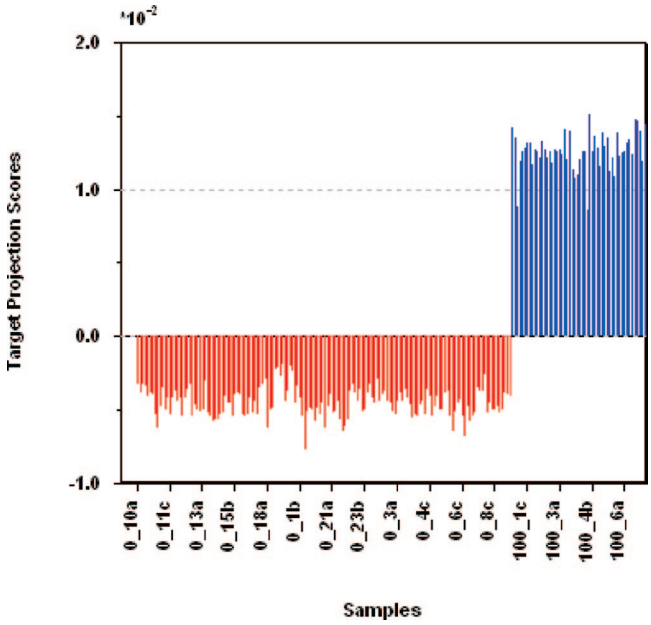

Figure 3. Scores on the target-projected (TP) component for reference samples (red) and samples spiked with 100 pM peptide standard (blue). PLS-DA model with 12 components transformed to produce the best possible single-component predictive model.

co-workers[25,26] using a window size of 20. Transformation from heteroscedastic to homoscedastic noise was carried out by a square root transform.[24,27] Normalization was performed to unit length. Representative spectra from the low mass range are displayed in Figure 1. All models were validated using cross-validation and permutation testing[28,29] on the response.

Software. The MALDI-TOF sampling and preanalysis (baseline correction) were performed using FlexAnalysis. Sirius version 8.0 from Pattern Recognition Systems was used for all additional analysis.

RESULTS AND DISCUSSION

Multivariate Analysis of Mass Spectral Profiles. Parts A and B of Figure 1 show spectra from a reference CSF sample and one sample with 100 pM peptide standard. There are differences between these spectra, but as shown in the following analysis, these differences are not due to the added multicomponent peptide standard but reflect natural variation in CSF composition in humans.

In order to remove outlying spectra from the data, PCA was performed for each group of samples, i.e., reference samples and samples spiked with 50, 100, 200, or 400 pM peptide standard. Six principal components were used for outlier detection, accounting for 65−70% of the variance in the data sets. Hotelling's T^2 and plots of residual standard deviation versus leverage were used to detect outliers. A total of 5−10% of the spectra showed features that made it necessary to remove them as outliers.

(25) Wong, J. W. H.; Durante, C.; Cartwright, H. M. *Anal. Chem.* **2005**, *77*, 5655–5661.
(26) Wong, J. W. H.; Cagney, G.; Cartwright, H. M. *Bioinformatics* **2005**, *21*, 2088–2090.
(27) Kvalheim, O. M.; Brakstad, F.; Liang, Y.-Z. *Anal. Chem.* **1994**, *66*, 43–51.
(28) Van der Voet, H. *Chemom. Intell. Lab. Syst.* **1994**, *25*, 313–323.
(29) Smit, S.; van Breemen, M. J.; Hoefsloot, H. C. J.; Smilde, A. K.; Aerts, J. M. F. G.; de Koster, C. G. *Anal. Chim. Acta* **2007**, *592*, 210–217.

After removal of outliers, the remaining spectra from each group of spiked samples were combined with the spectra from the group of reference samples. The number of samples differs slightly from group to group due to different number of analyzed samples and outliers removed (Table 2). For each group of spiked samples, a PLS-DA model was calculated with the group belonging as the response: zero (0) for the reference samples and one (1) for the spiked samples. Cross-validation was used to determine the number of significant PLS components in each model. The validation was performed by randomly selecting 50% of the samples as external validation set and repeating the analysis 10 times. The optimal number of components varied from run to run in the range of 10−13 components. We decided to use 12 components for all models. This choice gave explained variance in the range of 71.8−73.1% for the mass spectral data and in the range of 97.6−98.2% for the response discriminating reference samples from spiked samples (Table 2). The 12 component models were finally validated by permutation testing.[28,29] For each of the four PLS-DA models, target projection was performed to obtain a one-component predictive model. The target component represents the axis of optimal discrimination in the multivariate space spanned by the PLS model. Thus, the variance orthogonal to the response is removed and interpretation can be obtained on a single component. The variance in mass spectral profiles explained by the target component varies in the range of 2.2−4.8% for the four PLS-DA models (Table 2).

Analysis of Spectra from Reference Samples and Samples Spiked with 100 pM Peptide Standard. The results for the analysis of the combination of reference samples and samples spiked with 100 pM peptide standard are now presented. The two groups are not at all separated on any pair of extracted PLS-DA components, and the result for the two dominant components is shown in Figure 2. Figure 3 shows the scores on the target component. Excellent separation is observed. The reference samples are all negatively correlated to the spiked samples. The question is which of the variables are responsible for this separation and at which probability level? In order to answer this question, the SR and the percent CCR were calculated for all the 8881 binned m/z numbers. Completely random classification of the two groups on a variable corresponds to 50% CCR with equal number of samples in each group. This means that the reference and spiked samples have values ranging from the lowest to highest value on that variable and that the samples are randomly located on that variable. Such variables have no discriminatory ability at all. On the other hand, if a variable separates the two groups completely, all the samples of one group have lower values on that variable than the lowest value of that variable in the other group. Such a variable is on the top of the performance ladder with an ability to completely separate the two groups.

The 8881 binned m/z numbers were sorted according to their SR values. Mean correct classification rates and the standard deviation of MCCR were calculated for SR intervals of 0.1 from 0 up to 1.0. For SR higher than 1.0 the interval was increased to 0.25. Figure 4 shows MCCR plotted versus SR; i.e., the DIVA plot. In addition to probability for correct classification, boundaries corresponding to one standard deviation are also shown in the DIVA plot. Not surprisingly, MCCR increases almost monotonously with increasing SR until it levels out in intervals with

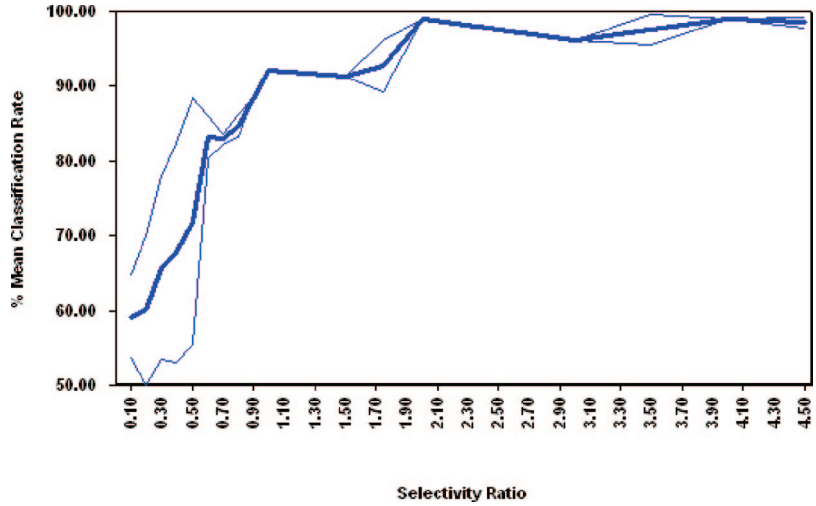

Figure 4. Discriminating variable (DIVA) plot for the TP model calculated using the reference samples and the samples spiked with 100 pM peptide standard: percent mean correct classification rate (MCCR) (thick line) and standard deviation of MCCR (thin lines).

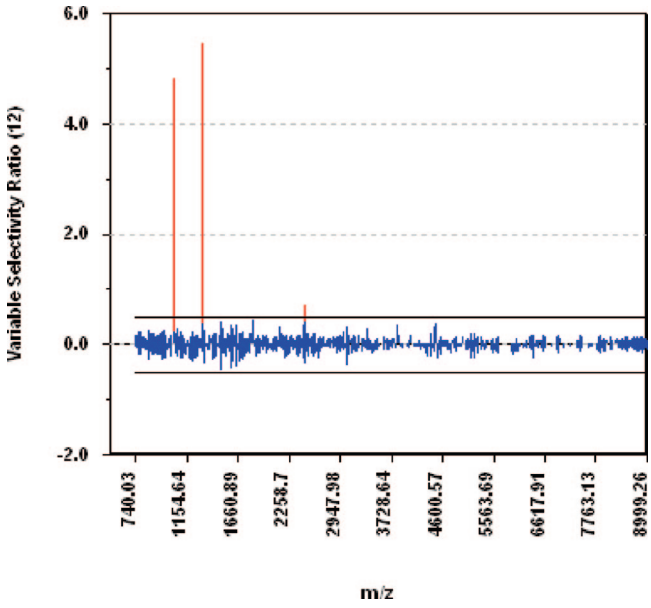

Figure 5. Selectivity ratio (SR) plot for the TP model calculated using reference samples and the samples spiked with 100 pM peptide standard. SR = 0.5 is marked by horizontal lines. The *m*/*z* numbers with a selectivity ratio exceeding this limit are marked in red. These are biomarker candidates at 80−85% MCCR, i.e., $p = 0.15-0.2$.

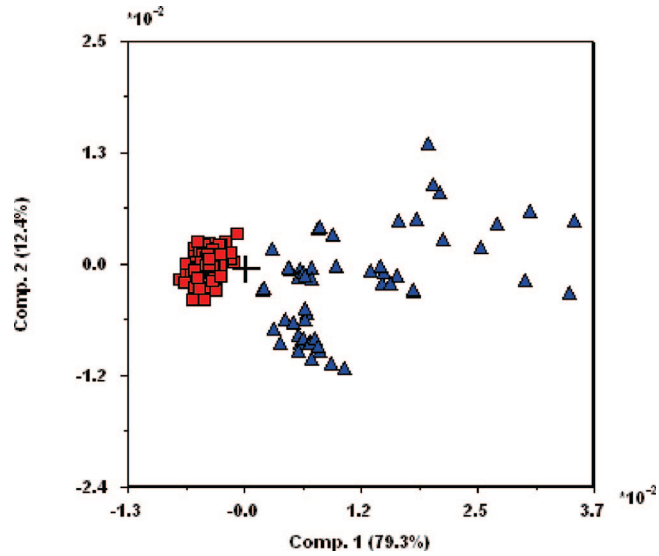

Figure 6. PCA score plot after variable selection using a probability level corresponding to $p = 0.15-0.2$ (SR = 0.5). Less than 0.3% of the original number of variables (i.e., 25 binned *m*/*z* numbers) is selected. Reference samples are in red, and spiked samples are in blue.

relatively high SR, i.e., variables with high between-to-within group variance and thus good discriminatory ability. Since MCCR corresponds to the mean probability of correct classification of the samples for each SR interval, we observe that variables with SR above 0.5 provide discrimination at a probability exceeding 80%, whereas MCCR converges toward 99% when SR approaches 2.0. Note also that standard deviation of MCCR gradually decreases with increasing SR values as expected since differences in misclassification decrease with increasing SR. For high SR, some intervals may have zero values since there may be no variables with SR in that range. In such cases, we interpolate between points and the standard deviation is set to zero. Figure 4 shows two such regions, one with SR in the interval of 1.0−1.5, and one with SR in the interval of 2.0−3.0.

Discriminating Variables and SR Plot. The relationship between SR and MCCR, and the interpretation that MCCR measures the discriminatory ability of the variables in a certain SR interval, provide the possibility of introducing probability-based boundaries in the SR plot. Figure 5 shows the SR plot with boundaries for the 8881 binned *m*/*z* numbers in the mass spectral profiles used to characterize the low MW fraction of the CSF samples. We have multiplied the selectivity ratios with the sign of the regression coefficients of the PLS-DA/TP model in order to make visible which variables are larger or smaller in the two groups of samples. Two regions of *m*/*z* numbers are revealed with almost perfect discriminatory ability between the reference samples and spiked samples. These two regions have SR in the range of 4.8−5.5 and are located in *m*/*z* regions 1048 and 1298. From Table 1 these signals are identified as originating from the peptides angiotensin II and angiotensin I. Just around SR = 0.75 (corresponding to 85% MCCR), we find a region corresponding

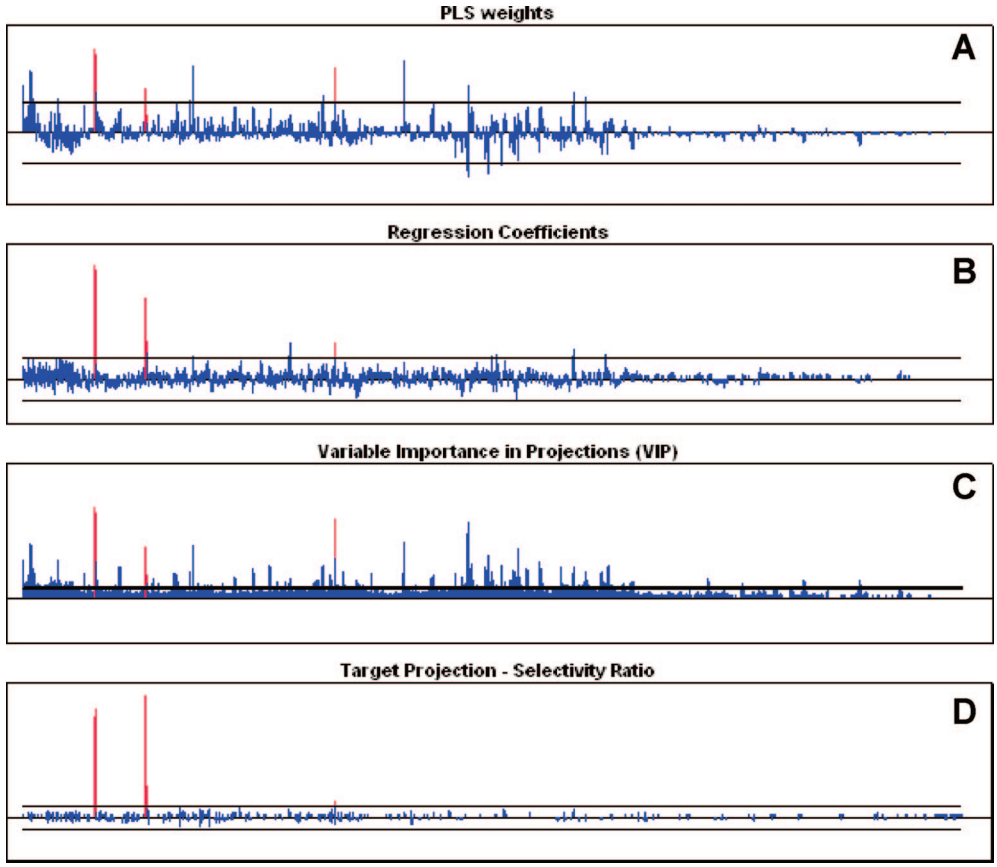

Figure 7. Comparison between different variable selection plots: (A) covariances between spectral variables and group belongings (PLS *X*-weights), (B) regression coefficients, (C) VIP (variable importance in projection) plot, and (D) selectivity ratio plot from target projection.

to m/z ratio of 2468. Table 1 shows that this signal corresponds to ACTH clip 18−39. The other peptides in the added standard are not visible above the threshold corresponding to SR = 0.5, i.e., 80−85% MCCR.

Variable Selection Using the DIVA Test and SR Plot. We can use the probability distribution obtained from the DIVA test for variables selection. Figure 6 shows the PCA score plot for variable selected according to a probability level corresponding to $p = 0.15-0.2$ (SR = 0.5). Excellent separation is obtained with only 25 binned m/z numbers, i.e., less than 0.3% of the original number of variables. The larger spread of spiked samples reflects the experimental variation resulting from spiking with small volumes of samples available. If the variable selection is performed in order to select biomarkers for further investigation, the investigator has to decide what probability level to use. A choice of, e.g., $p = 0.1$ compared to, e.g., $p = 0.25$, means that the risk of assigning false biomarkers is decreased on the expense of the possibility of leaving out some real biomarkers. The choice may depend on other factors than just statistical significance. For instance, a criterion based on the actual amount of work to further investigate the selected biomarkers may be of practical concern, and then a limit providing fewer rather than more biomarker candidates may be more practical.

Comparison with Other Methods for Variable Selection. As discussed in the introduction, numerous methods exist for variable selection. The methods differ in aims as well as performance. In the following, we compare the result of our selection with the result of three other methods for variable selection that are commonly used by PLS-DA practitioners and also implemented in many of the available software packages. The three methods are PLS weights,[25] size of regression coefficients,[26] and VIP.[27] The results for these methods together with the SR plot are displayed in Figure 7. For each method, we have defined thresholds of selections. For PLS weights and regressions coefficients, we select the variables exceeding two standard deviations around the mean. This choice, which is rather conservative, still leads to too many false biomarker candidates. For VIP, Figure 7C shows that the recommended threshold of 1.0[27] for selection leads to a forest of false biomarker candidates. If we increase the threshold, the number of false candidates is reduced, but we also start to lose real biomarkers. The SR plot with boundaries corresponding to 80% MCCR (SR = 0.5) provides only correct candidates.

One may argue that even if too many variable regions are selected, this may not influence the classification results. In order to assess this possibility, we have used PCA on the selected regions for the model with reference samples and samples spiked with 100 pM peptide standard. The result is shown in Figure 8. By using retaining variables for which PLS weights exceed two standard deviations from the mean, 90 variables were selected. We observe strong overlap between groups on the two dominant principal components which explain almost two-thirds of the variation in selected variables. The result of using absolute size of regression coefficients as selection criterion is much better (Figure 8B). A total of 39 variables were selected providing a separation into two groups. If we compare the classification with the results from SR plot (Figure 6), however, we note that the group of controls has spread out. This is a consequence of

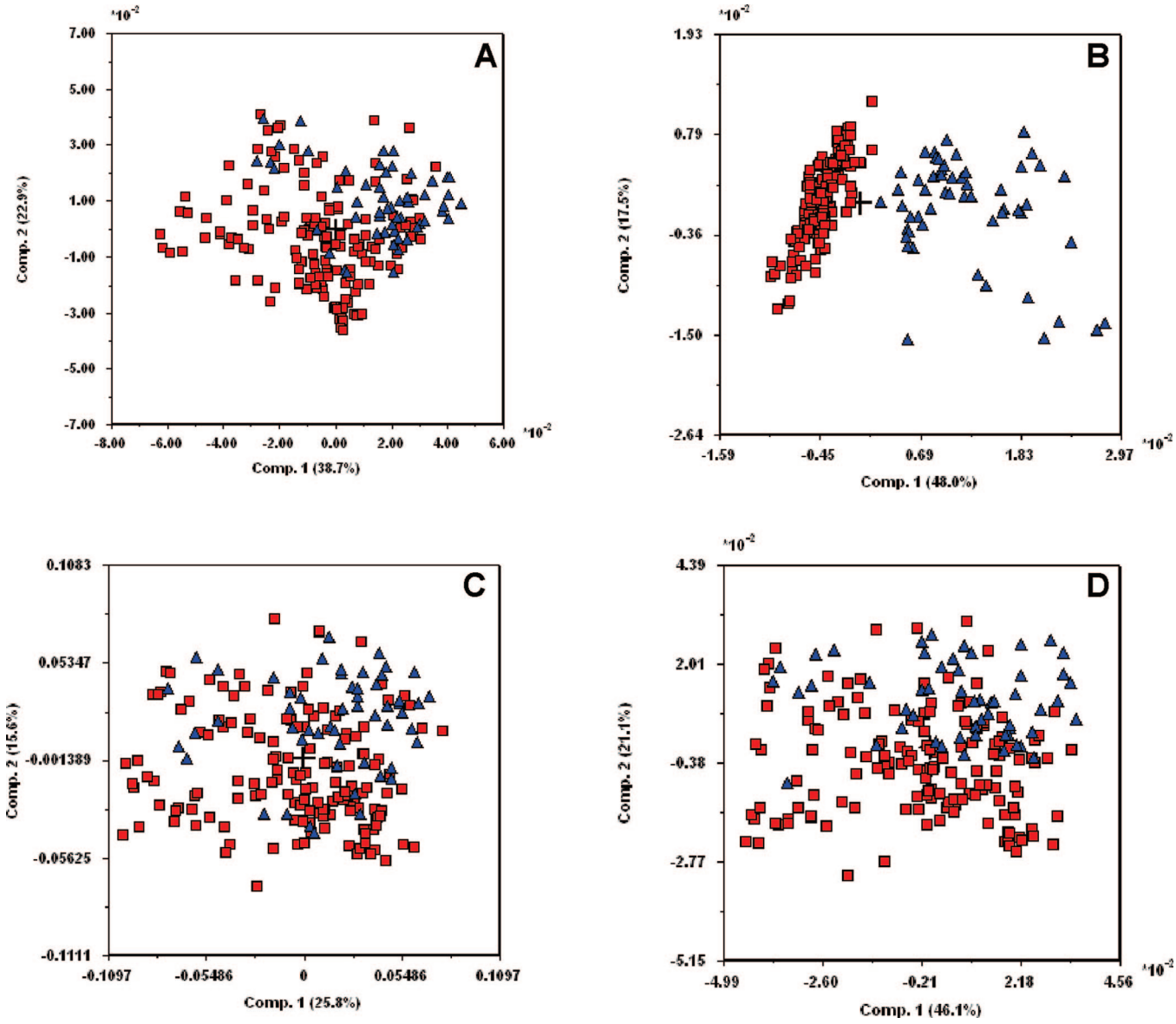

Figure 8. PCA score plot after variable selection using (A) covariances between spectral variables and group belongings (PLS *X*-weights), (B) regression coefficients, (C) VIP (variable importance in projection) plot with recommended limit of 1.0 for selection, and (D) VIP plot with a limit of 5.0 for selection. Reference samples are in red, and spiked samples are in blue.

inclusion of orthogonal variation, i.e., variables with large regression coefficients unrelated to the response have been incorporated. The selection using VIP with threshold 1.0 results in 1615 variables, and complete overlap of groups is observed (Figure 8C). If we increase the threshold to 5.0 and redo the selection, some tendency of grouping is observed (Figure 8D), but still orthogonal variation impacts the results and, in addition, real biomarkers are lost in the selection process. We conclude that our novel approach is superior to well-established methods for variable selection in PLS-DA.

Results for the Other Models. The same analysis as above for reference and group with spiked with 100 pM standard was performed for the other three models. The results were similar and are therefore not shown here. However, for the model with samples spiked with 400 pM peptide standard, a variable with SR ≈ 1 was observed around m/z 3620. This would correspond to a false biomarker candidate and is to be expected with the number of variables outsizing the number of samples with a factor of 50. Furthermore, for the model with samples spiked with 50 pM peptide standard, the highest SR is 1.5 and the SR for ACTH clip 18−39 falls below the probability level for being identified as an important discriminating variable. These observations result from the depletion effect of spiking with smaller and smaller amounts of peptide standard.

Comparison of the DIVA Distribution for All Models. We can compare the results for the four models by plotting MCCR versus SR. Figure 9 shows this plot based on the models obtained from combining reference samples with samples spiked with 50, 100, 200, or 400 pM peptide standard. There are no selectivity ratios higher than 1.5 for the 50 pM model due to the small amount of added peptide standard. Otherwise, the probability distribution is strikingly similar for all models showing that for the same type of instrumental technique and same type of samples the nonparametric DIVA test appears to be reasonably robust. This observation indicates that it may be possible to come up with approximate estimates of significance levels to be used in the SR

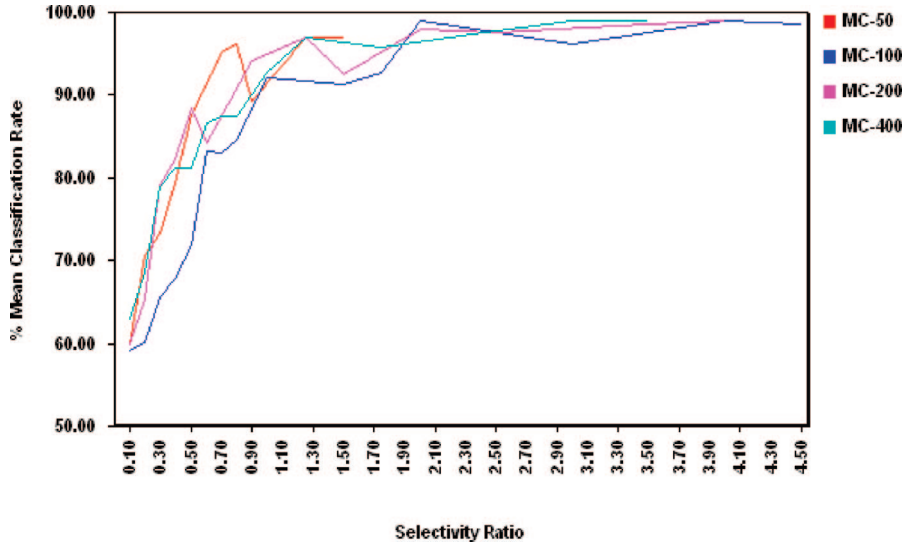

Figure 9. Percent mean correct classification rate (MCCR) for the TP models calculated from the reference samples and the samples spiked with 50 (red), 100 (blue), 200 (lilac), or 400 pM (turquoise) peptide standard.

plot without the need to run the DIVA test for each new data set in similar applications.

CONCLUSIONS

Numerous methods for variable selection in multivariate classification and regression problems have been devised. In this work, we have designed a nonparametric test that provides a probability measure to guide selection of variables (e.g., biomarkers) from complex multicomponent profiles ensure variables with both good explanatory and good predictive performance.

The probability level quantifies the discriminatory ability of a variable and, in a binary classification, corresponds to a mean sensitivity for the variables within the same selectivity ratio interval. This mean sensitivity is strongly correlated to the selectivity ratio, a property that furnishes quantitative limits for the discriminatory ability of variables. By using the information of negative and positive correlations between variables in two groups of samples the selectivity ratio plot displays which variables are increasing or decreasing from one group of samples to the other. Thus, in biomarker applications, the SR plot shows which candidates are up or down regulated.

The SR plot with probability boundaries obtained from the nonparametric DIVA test provides the investigator with an excellent objective tool to assist the selection of discriminating variables/biomarker candidates for further exploration. Selecting variables according to a low limit for the selectivity ratio increases the risk of selecting false candidates, whereas a high limit for the selectivity ratio increases the risk of loosing potential candidates. Due to the continuous nature of the distribution of the probabilities available from the DIVA test, the investigator can make a rational choice for the limit between interesting and less interesting variables to fit his particular application and the resources and possibilities he has for further investigations.

ACKNOWLEDGMENT

This work was supported by Grants from the aid of EXTRA funds from the Norwegian Foundation for Health and Rehabilitation, the Kjell Almes Legacy, the Bergen MS Society, the Odd Fellow, the Meltzer Foundation, the Norwegian Society of Multiple Sclerosis, and Fritz and Ingrid Nilsen's Legacy for Research in Multiple Sclerosis, Norway. The authors are partly supported by the National Programme for Research in Functional Genomics (FUGE), funded by the Norwegian Research Council, and the Western Norway Regional Health Authority. Pattern Recognition Systems AS is thanked for partial financing of R. Arneberg and for providing the Sirius software free of charge. Professor Rune Johan Ulvik, Professor Christian A. Vedeler, and Ph.D. Frode Berven are acknowledged for support. Two anonymous reviewers are thanked for valuable comments.

Received for review November 27, 2008. Accepted January 28, 2009.

AC802514Y