Towards the Potential use of ^{1}H NMR Spectroscopy in Urine Samples for Prostate Cancer Detection

Patricia Zaragozá,[a] Jose Luis Ruiz-Cerdá,[b] Guillermo Quintás,[c] Salvador Gil, [a,d] Zacarías León,[c] Jose Luis Vivancos,[a,e] and Ramón Martínez-Máñez[a,e]

Supporting Information

Sample Procedure

Normal spot urine samples from were collected in 2 mL disposable polyethylene containers. Urine samples for the detection of PCa were collected from prostate cancer patients, at La Fe Hospital, Valencia, Spain. The collected urine samples were frozen and stored at −80 C until analyses. Samples were centrifuged at 2500 rpm for 5 minutes to eliminate solids and other insoluble material, and then aliquoted.

NMR Procedure

The reference solution was prepared dissolving 25,2 mg of 4,4-dimethyl-4-silapentane-1-sulfonic acid (DSS) in deuterated water (D_2O, 5,758 g). Each sample was prepared from the corresponding raw urine sample (450 uL) by addition of 40 uL of the Reference solution.

From each sample three proton nmr spectra (at 310 K) were recorded: a standard proton nmr, a proton nmr with presaturation sequence on the water signal (main set used in the study) and a 1D-diffusion sequence.

The standard proton allows ensuring the chemical shift of the different signals, which can be slightly changed by the presaturation sequence.

1D-Difusion sequence was performed to enhance signals from high molecular weight substances and at the same time reducing the signal from the water, and from low molecular weight substances.

A study of the line-width of the reference compound DSS in all spectra has been carried out. A box plot has been used to illustrate the numerical data obtained. Figure S1 shows the deviation of the line-width of the reference compound DSS in the ^{1}H NMR spectra of the samples.

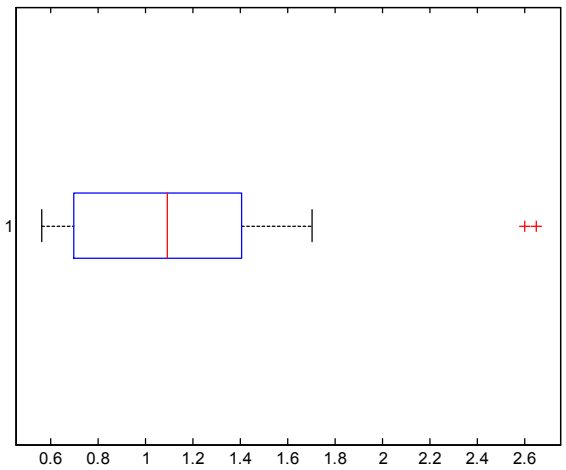

Figure S1. Box plot of the line-width of the reference compound DSS in the ^{1}H NMR spectra.

Data Analysis

Data analysis was performed using MATLAB 2012b (The Matwhorks), the PLS Toolbox 7.0 (Eigenvector Res. Inc.), the icoshift function available from www.models.life.ku.dk and in house written MATLAB scripts.

NMR spectra acquired were imported into MATLAB® (2012b, The Mathworks Inc. Natick, MA, USA). The interval correlation shifting (icoshift) algorithm developed by Savorani et al.[S1] was used for initial spectral alignment to overcome shifts of pH dependent signals found in the data set. The icoshift algorithm alignes each NMR feature to a target (in this work, the median spectrum of the whole spectral data set) by maximizing the cross correlation between user defined intervals. Here, the NMR spectra were split into 51 intervals selected after visual inspection of the regions according to common spectral features among samples.

Data (NMR signal) was (column-wise) centered and scaled using the mean and standard deviation of each variable in the calibration set, respectively. Centering adjusts for differences in the offsets between high and low intensity signals at different chemical shifts leaving only the variation between samples for analysis. After mean centering, each variable was scaled using its standard deviation as scaling factor (i.e. autoscaling, also known as unit variance scaling) thus allowing data analysis on the basis of correlations instead of covariances. No further normalization factors were used.

In Prediction Results, it is possible to define Sensitivity (S_n) as a measure of the fraction of the predicted sites that are correct amongst those predicted: $S_n = T_p/(T_p + F_n)$ where T_p are True Positives and F_n are False Negatives. And it is also possible to define Specificity (S_p) as a measure of the fraction of the predicted that are correct amongst those predicted: $Sp = T_p/(T_p + F_p)$ where T_p are True Positives and F_p are False Positives [S2].

Samples Analyzed

A total of 113 samples were used and split into calibration and validation subsets. As control (without PCa), patients after radical prostatectomy and patients diagnosed benign prostatic hyperplasia (BPH) were used. The randomly selected calibration subset included a total of 49 samples collected from 21 patients with PCa and a total of 28 control samples (17 after radical prostatectomy and 11 diagnosed with BPH). The validation set was formed by 64 samples including 50 PCa samples, and 14 samples classified as control (9 after radical prostatectomy and 5 diagnosed with BPH).

PCA

Initially, a principal component analysis (PCA) model was built using the calibration set and autoscaling as data pretreatment. From the scores plot of the first versus the second principal components and from the Q-residual values shown in the Figure S2 no samples included on the calibration set were classified as outliers.

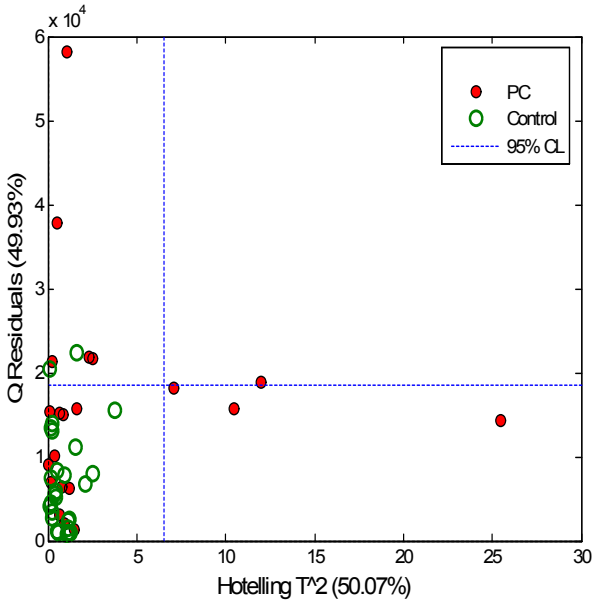

Figure S2. Q residual and Hotelling-T^2 values calculated using a two PC model calculated using the calibration data set and autoscaling as pretreatment.

PLSDA Model

Supervised discriminant analysis was performed using partial least squares (PLSDA) and a maximum number of 5 latent variables (LVs). The X-block (i.e. NMR data) was autoscaled and the y vector containing class labels (i.e. -1 and +1 for control and PC samples, respectively) was mean centered. The residual Q and the Hotelling's T^2 statistics were also used for outlier detection. Selection of the number of Latent Variables using Leave-one-out crossvalidation: LV=min{NMC=FP+FN} (Figure S3).

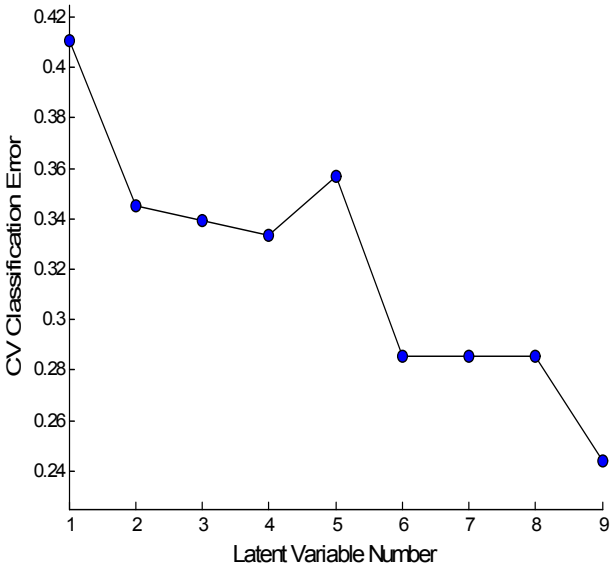

Figure S3. Number of latent variables vs the error of classification using leave-one-out crossvalidation.

Initial PLSDA figures of merit were obtained by after 11 iterations of a random 5-fold cross validation. From cross validation data, 3 latent variables were retained.

Then, a selection of the most differentiating spectral features was carried out based on the variable importance scores vector (VIP) calculated from the initial PLSDA model. Since the average of the squared VIP scores equals 1, the greater than 1 criteria is used as a rule of thumb for variable elimination[S3]

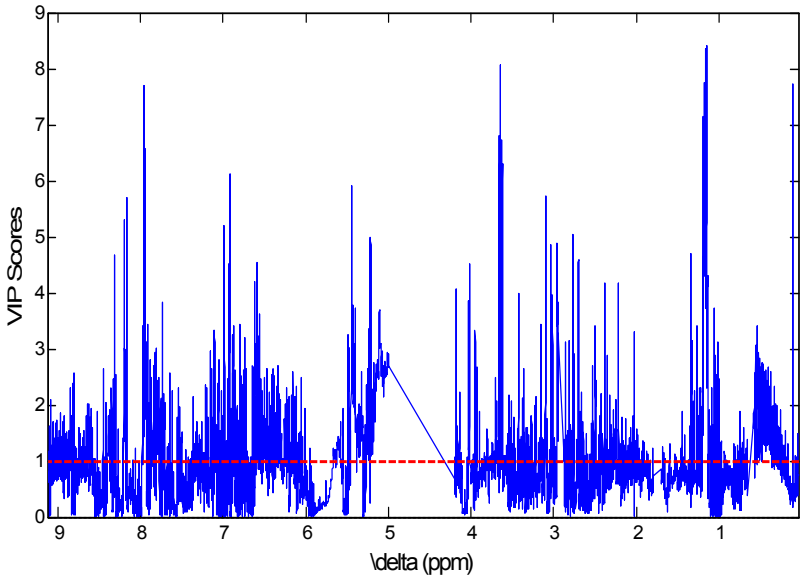

Figure S4. VIP Scores of the PLSDA using 2 latent variables.

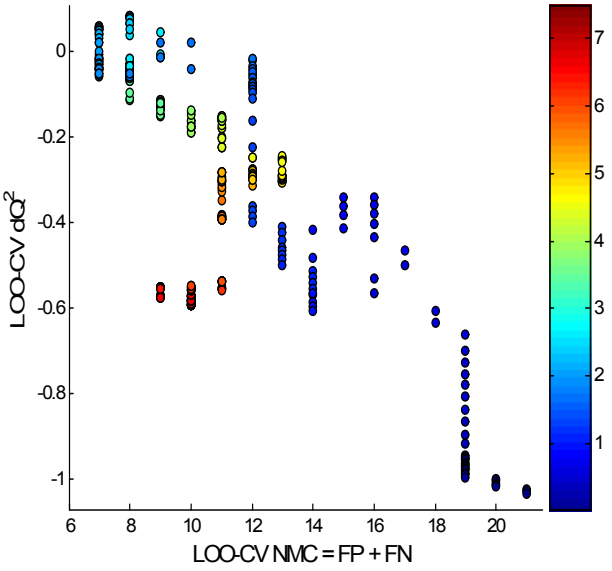

Figure S5. dQ^2 vs number of samples misclassified using Leave-one-out cross validation.

PLSDA Model using VIP cutoff=2.28

Instead, the effect of using VIP cutoff values in the 0-7 range was evaluated by leave one out cross validation using the discriminant Q^2 (dQ^2) statistic and the number of misclassified (NMC) samples

as target. Based on a VIP cutoff value of 2.28 a total of 1627 variables were retained and used for the calculation of a second PLSDA model (Figure S6).

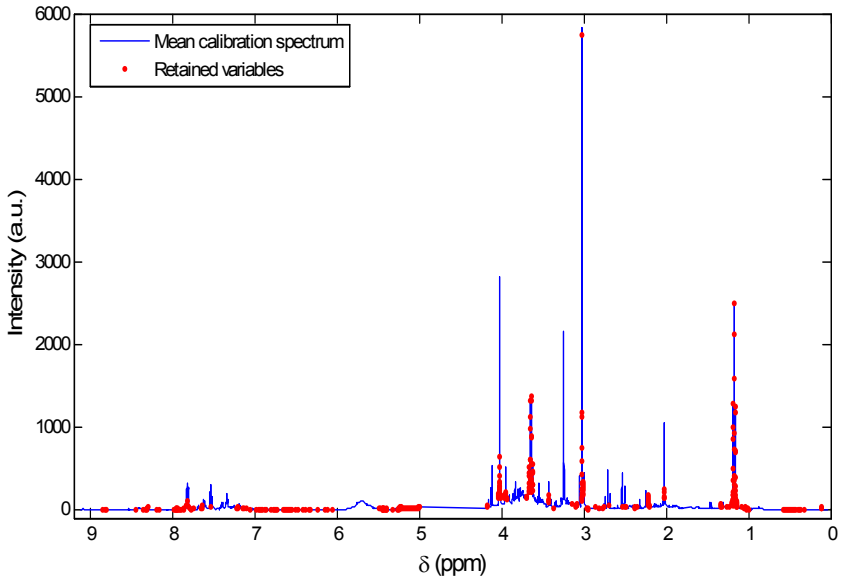

Figure S6. Retained variables using a VIP cutoff=2.28 (1627 variables)

Supervised discriminant analysis was performed using partial least squares (PLSDA) and a maximum number of 5 latent variables (LVs). The X-block (i.e. NMR data) was autoscaled and the y vector containing class labels (i.e. -1 and +1 for control and PC samples, respectively) was mean centered. The residual Q and the Hotelling's T^2 statistics were also used for outlier detection. Selection of the number of Latent Variables using Leave-one-out crossvalidation: LV=min{NMC=FP+FN} (Figure S7).

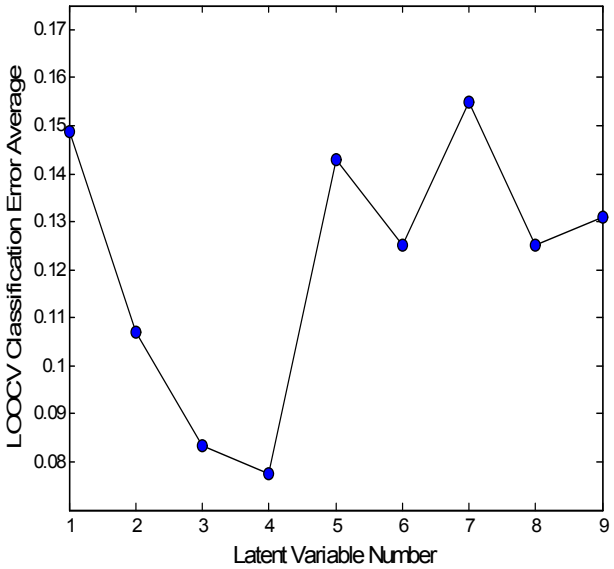

Figure S7. Number of latent variables vs the error of classification using leave-one-out crossvalidation.

PLSDA Model using metabolites in a previous work

The high spectral overlapping did not allow the use of specific signals of metabolites for the development of univariate or multivariate models for PCa discrimination. However, we developed a PLSDA model using the [0.9-1.55; 1.68-1.88; 2-2.82; 3.1-3.34; 3.97-4.17] (ppm) interval (see Figure S8) covering a set of metabolites that have been shown to be related to PCa in previous work.[S4] The set of metabolites included: myo-inositol (4.07 ppm), phosphocholine (3.24 ppm), spermine(1) (3.2 ppm), citrate(1) (2.72 ppm), citrate(2) (2.58 ppm), glutamine (2.36 ppm), spermine(2) (2.11 ppm), spermine(3) (1.78 ppm), alanine (1.45 ppm), lactate (1.32 ppm), OH-butyrate (1.19 ppm) and Valine-Leucine (1.01 ppm).

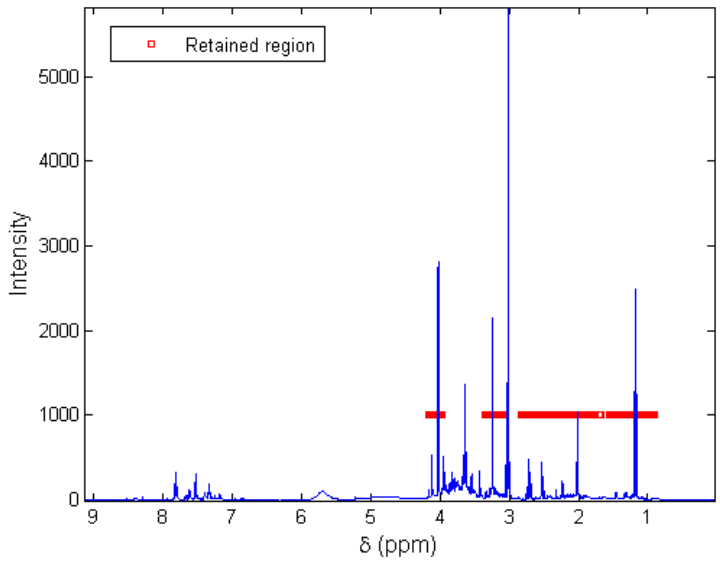

Figure S8. Interval region (red) selected for the calculation of a PLSDA model.

Using the selected variable interval, a PLSDA model was developed. Briefly, an initial PLSDA model was developed using the 49 calibration samples and the selected variable intervals. Then, 1555 variables showing VIP scores values >1 in the PLS model were selected for the development of a second PLSDA model. The predictive performance of the second PLSDA model was evaluated using the external validation set comprising 64 samples.

Figure S9 shows the predicted values obtained using 5488 variables included in the intervals shown in Figure S8. The statistical significance of leave-one-out-CV (LOOCV)-error was estimated by permutation testing (num. permutations=500). Results summarized in the table below showed that the model performance was not statistically significant (p-value=).

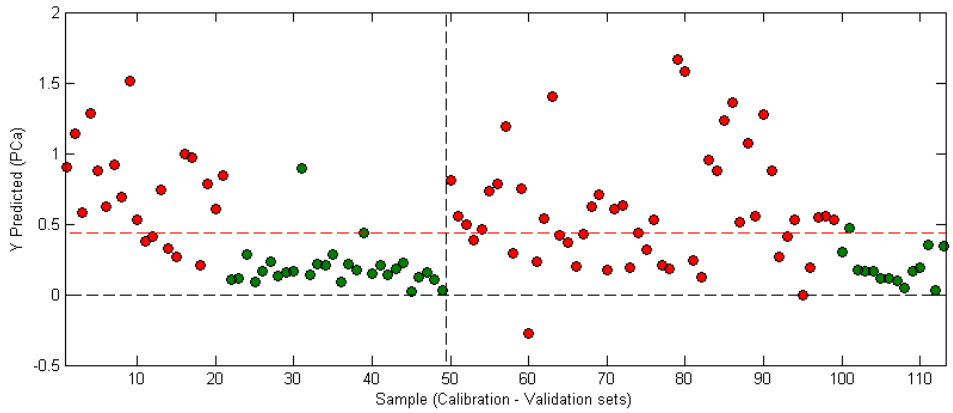

Figure S9. Predicted values by a PLSDA model calculated using the selected spectral interval depicted in Figure S8.

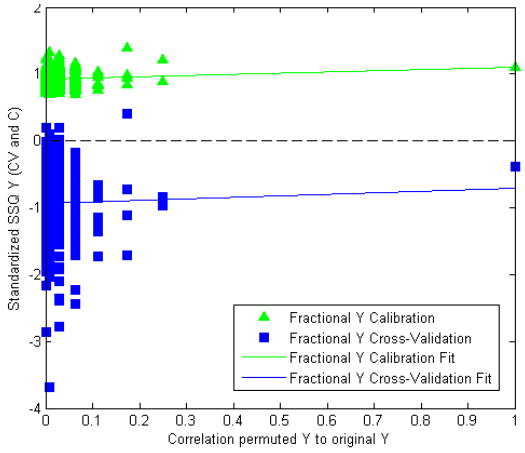

Figure S10. Results from the permutation test of a PLSDA model calculated using the selected spectral interval depicted in Figure S8. (LVs=2)

Table S1. Probability of Model Insignificance vs. Permuted Samples.

	Wilcoxon	Sign test	Random t-test
Self prediction	0.074	0.117	0.535
LOO-CV	0.084	0.091	0.421

Figure S11 shows the predicted y-values using a PLSDA model calculated based on 1555 variables with VIP>1 in the initial PLSDA model. Results obtained in this model summarized in the confusion table included below were clearly worse than those included in the manuscript based on a variable selection using the whole interval range and the same feature selection procedure.

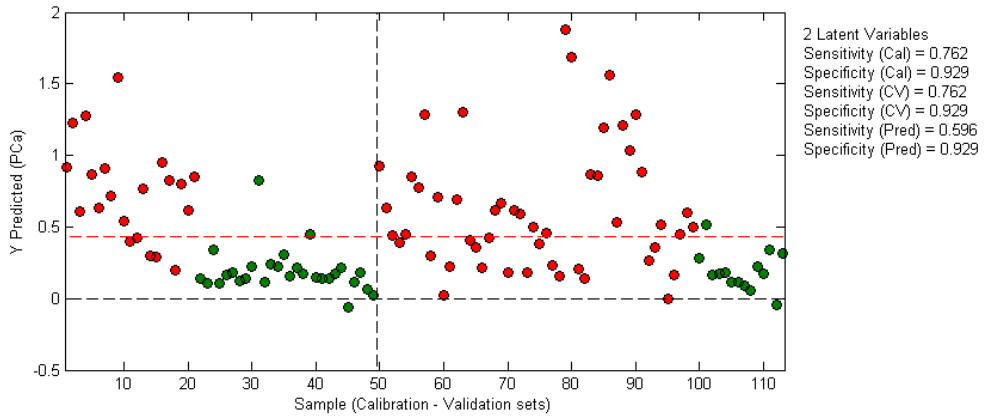

Figure S11. Predicted values by a PLSDA model calculated using a PLSDA model calculated based on 1555 variables with VIP>1 in the initial PLSDA model.

Table S2 Confusion table for the prediction set (i.e. external validation) after variable selection.

	PCa	Control
Predicted as PCa	31	1
Predicted as Control	19	13

References:

S1. F. Savorani, G. Tomasi and S.B. Engelsen, *Journal of Magnetic Resonance,* 2010, **202**, 190.

S2. D. J. Balding, M. Bishop and C. Cannings, *Handbook of Statistical Genetics,* Wiley, New York, 3rd edn., 2007.

S3. I.G. Chong and C.H. Jun, *Chemometrics and Intelligent Laboratory Systems*, 2005, **78**, 103.

S4: N.J. Serkova, E.J. Gamito, R.H. Jones, C. O'Donnell, J.L. Brown, S. Green, H. Sullivan, T. Hedlund and E.D. Crawford, *Prostate,* 2008, **68**, 620.

Matlab Script

PLSDA Model using 21934 variables and Leave-one-out cross validation selecting 2 Latent Variables

```
% Cal X: NMR_BLCmed_calib (49x21934)
% Val X: NMR_BLCmed_valid (64x21934)
% variables included = include_PLSDA vector
% PLSDA model
% LV=2 (LOO-CV)
% X: scaling = autoscale
% Y: mean centering
% Model details
% Statistics for each y-block column:
% Modeled Class: 2   3
% Sensitivity (Cal):  0.810  0.964
% Specificity (Cal):  0.964  0.810
% Class. Err (Cal): 0.113095 0.113095
% RMSEC: 0.340608 0.340608
% Bias:     0 -1.11022e-016
% R^2 Cal: 0.526277 0.526277
%
%    Percent Variance Captured by Regression Model
%
%           -----X-Block-----    -----Y-Block-----
%   Comp     This      Total      This      Total
%   ----    -------   -------    -------   -------
%     1      33.78     33.78      23.85     23.85
%     2       7.50     41.27      28.78     52.63
% save model: plsdamodel_Cal_LV2
% Confusion Matrix:
%    Class:      TP          FP          TN          FN
%      PRE       0.80952     0.03571     0.96429     0.19048
%      POST      0.96429     0.19048     0.80952     0.03571
%
% Confusion Table:
%                                Actual Class
%                           PRE      POST
% Predicted as PRE          17        1
% Predicted as POST          4       27
%
% CV RESULTS
% Confusion Matrix (CV):
%    Class:      TP          FP          TN          FN
%      PRE       0.52381     0.25000     0.75000     0.47619
%      POST      0.75000     0.47619     0.52381     0.25000
%
% Confusion Table (CV):
%                                Actual Class
%                           PRE      POST
% Predicted as PRE          11        7
% Predicted as POST         10       21
%
% LOO-CV errors were tested using a Permuation test (nperm=500)

VIP_all=vip(plsdamodel_Cal_LV2);
VIP_all(:,2)=[];
vselect=find(VIP_all>=2.28);
% Note: VIP cutoff selection based on LOO-CV results at different VIP
% values.
```

PLSDA Model using VIP cutoff=2.28 and 1627 variables and Leave-one-out cross validation selecting 2 Latent Variables

```
% New PLSDA model using only the retained variables
% Cal X: NMR_BLCmed_calib (49 x vselect) (49x1627)
% Val X: NMR_BLCmed_valid (64 x vselect) (64x1627)
% variables included = include_PLSDA vector
% PLSDA model
% LV=2 (LOO-CV)
% X: scaling = autoscale
% Y: mean centering
% Model details
% Num. LVs: 2
% Cross validation: leave one out
% Statistics for each y-block column:
% Modeled Class: 2   3
% Sensitivity (Cal):  0.905  0.929
% Specificity (Cal):  0.929  0.905
% Sensitivity (CV):  0.857  0.929
% Specificity (CV):  0.929  0.857
% Class. Err (Cal): 0.0833333 0.0833333
% Class. Err (CV): 0.107143 0.107143
% RMSEC: 0.311563 0.311563
% RMSECV: 0.371116 0.371116
% Bias: -5.55112e-017 1.11022e-016
% CV Bias: -0.000251722 0.000251722
% R^2 Cal: 0.603624 0.603624
% R^2 CV: 0.455276 0.455276
%
%    Percent Variance Captured by Regression Model
%
%           -----X-Block-----    -----Y-Block-----
%   Comp     This      Total      This      Total
%   ----    -------   -------    -------   -------
%     1      22.46     22.46      49.65     49.65
%     2      27.71     50.17      10.71     60.36
%
% Figure scores plot / Predicted values / ....
% MODEL RESULTS
%
% MODEL RESULTS
%Confusion Matrix:
%    Class:      TP          FP          TN          FN
%      PRE       0.90476     0.07143     0.92857     0.09524
%      POST      0.92857     0.09524     0.90476     0.07143
%
%Confusion Table:
%                                Actual Class
%                           PRE       POST
%Predicted as PRE          19         2
%Predicted as POST          2        26
%
%CV RESULTS
%Confusion Matrix (CV):
%    Class:      TP          FP          TN          FN
%      PRE       0.85714     0.07143     0.92857     0.14286
%      POST      0.92857     0.14286     0.85714     0.07143
%
%Confusion Table (CV):
%                                Actual Class
%                           PRE       POST
%Predicted as PRE          18         2
%Predicted as POST          3        26
%
%
%PREDICTION RESULTS
%Confusion Matrix:
%    Class:      TP          FP          TN          FN
%      PRE       0.72000     0.00000     1.00000     0.28000
%      POST      1.00000     0.28000     0.72000     0.00000
%
%Confusion Table:
%                                Actual Class
```

```
%                             PRE       POST
% Predicted as PRE           36         0
% Predicted as POST          14        14
%
% 2 Latent Variables
% Sensitivity (Cal) = 0.905
% Specificity (Cal) = 0.929
% Sensitivity (CV) = 0.857
% Specificity (CV) = 0.929
% Sensitivity (Pred) = 0.720
% Specificity (Pred) = 1.000
```