Classification of samples from NMR-based metabolomics using principal components analysis and partial least squares with uncertainty estimation

Werickson Fortunato de Carvalho Rocha,[1] David Sheen,[2] and Daniel W. Bearden[3]

[1]National Institute of Metrology, Quality and Technology -INMETRO, Division of Chemical Metrology, 25250-020 Duque de Caxias, RJ, Brazil

[2]Chemical Sciences Division, National Institute of Standards and Technology, Gaithersburg, MD 20899, USA

[3]Chemical Sciences Division, National Institute of Standards and Technology, Hollings Marine Laboratory, 331 Fort Johnson Road, Charleston, SC 29412, USA

Corresponding author: David Sheen

ORCID: 0000-0003-1958-1848

david.sheen@nist.gov

Acknowledgements

This work was partially supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (National Council for Scientific and Technological Development) of Brazil [grant number REF.203264/2014-26]. The authors also thank Dr. Katrice Lippa, Dr. David Duewer, and Dr. Pamela Chu at NIST for productive discussions.

Abstract

Recent progress in metabolomics has been aided by the development of analysis techniques such as gas and liquid chromatography coupled with mass spectrometry (GC-MS and LC-MS) and nuclear magnetic resonance (NMR) spectroscopy. The vast quantities of data produced by these techniques has resulted in an increase in the use of machine algorithms that can aid in the interpretation of this data, such as principal components analysis (PCA) and partial least squares (PLS). Techniques such as these can be applied to biomarker discovery, interlaboratory comparison, and clinical diagnoses. However, there is a lingering question whether the results of these studies can be applied to broader sets of clinical data, usually taken from different data sources. In this work, we address this question by creating a metabolomics workflow that combines a previously-published consensus analysis procedure (10.1016/j.chemolab.2016.12.010) with PCA and PLS models using uncertainty analysis based on bootstrapping. This workflow is applied to NMR data that come from an interlaboratory comparison study using synthetic and biologically-obtained metabolite mixtures. The consensus analysis identifies trusted laboratories, whose data are used to create classification models that are more reliable than without. With uncertainty analysis, the reliability of the classification can be rigorously quantified, both for data from the original set and from new data that the model is analyzing.

Keywords: Metabolomics; reliability; bootstrap; uncertainty estimation; chemometrics; biomarker discovery

1 Introduction

Metabolomics is an emerging field concerned with developing profiles of small molecule metabolites in biological systems. These profiles describe the different physiological states, including disease states, that a system may adopt. Progress in metabolomics has been aided by the advent of analysis technologies for comprehensive metabolic analysis [1]. Different analysis techniques have been applied to this problem, including liquid chromatography-mass spectrometry (LC-MS) [2-4], ambient ionization mass spectrometry [5-7], ultra-high performance liquid chromatography-two-dimensional mass spectrometry (UHPLC-MS/MS) [8-10], nuclear magnetic resonance (NMR) spectroscopy [4, 11-13], and Raman spectroscopy [14-16]. Due to this broad range of analytical techniques, an enormous volume of data is now available at various levels of complexity. Interpretation of such data has required the development of new data analysis procedures.

Principal components analysis (PCA) is perhaps the most widely used data analysis tools used in metabolomics [17, 18], partly due to the prevalence of computer packages that implement it and the unsupervised nature of the analysis. PCA results in a model that identifies features of the data with the greatest variance between measurements. Other common methods include supervised approaches such as linear discriminant analysis (LDA) [19, 20], partial least squares-discriminant analysis (PLS-DA) [21, 22], support vector machine for classification (SVM-DA) [23-25] and soft independent modeling by class analogy (SIMCA) [26]. These supervised methods result in models that identify features of the data that are correlated with distinction between classes.

Quality assurance and control in measurements is a crucial but overlooked component of the metabolomics pipeline [27]. Establishing data quality requirements across dispersed, cooperating laboratories is a difficult challenge, especially for non-targeted analysis. One approach is to

develop data quality metrics for spectral measurements which provide a fair assessment of the raw or minimally-processed data which does not rely, for example, on high-level processing such as feature selection or compound identification. In previous work [28], we proposed a technique for scoring NMR spectral data and determining trusted laboratories based on that scoring process. That workflow used a consensus-analysis technique to determine trusted laboratories based on their ability to generate reproducible data for simple systems and then used the data from the trusted laboratories on a more complex system as an input to a chemometric analysis technique, in this case PLS-DA with uncertainty analysis. We used a wrapper to perform a residual bootstrap analysis [29, 30] of a standard PLS-DA algorithm. The output is a model that provides an uncertainty estimate on all predictions that it makes.

In this work, we propose to make the previously developed scoring system part of a complete chemometric workflow for interlaboratory metabolomics studies. The objective of this work is to apply our previous work in uncertainty analysis for PCA and PLS-DA [31] to data obtained from a multi-laboratory intercomparison exercise for environmental metabolomics; however, the workflow is applicable to any properly designed metabolomics intercomparison study. This approach demonstrates the use of different statistical methods to estimate the uncertainty of the results obtained by chemometric models in multi-laboratory intercomparison exercises.

1.1 Quality assurance and quality control in chemometrics

For data to be broadly useful to the scientific community, that data must be reasonably reproducible by any scientist with a similar level of experience and similar instrumentation. A recent survey of metabolomics researchers [27] showed that there is a relatively haphazard implementation of widely-accepted methods for ensuring quality of measurements, such as formalized standard operating procedures (SOPs) or data validation procedures. Interlaboratory

studies with defined sample protocols are an important part of encouraging quality control [12, 32], whether the studies perform spectral feature assessments, pattern recognition assessments or quantitative assessments. Our previous analysis of an interlaboratory study [28] set out to score laboratories based on direct assessment of minimally-processed, binned spectra using measures of geometric closeness of the spectra, avoiding the need for expert evaluation of the data such as compound identification or quantification.

1.2 Uncertainty analysis in chemometrics

Analytical results are not complete unless they are expressed with the uncertainty in their predictions, meaning the range of values that can be reasonably attributed to an analytical result considering a defined probability of error (or a level of confidence) [33]. Traditionally, the performance of chemometrics models is assessed using statistical parameters based on the model prediction error, such as root-mean-squared error of calibration (RMSEC), root-mean-squared error of prediction (RMSEP), specificity, sensitivity, and number of misidentified samples. All of these metrics give a broad overview of the reliability of the model's predictions. What is needed, however, is a way to assess the prediction uncertainty of a particular unknown case, so that in a forensic or clinical environment it is possible to estimate the trustworthiness of any new measurement.

Many publications have explored sample-specific uncertainty and reliability analysis in one form or another [34-36], often using some variation on bootstrap analysis [37-43]. For instance, Faber and coworkers [44, 45] and Martens and Martens [46] analyzed uncertainty in PLS regression using bootstrap and jackknifing. The work of Wentzell and co-workers has probed the uncertainty structure of NMR metabolomics data similar to the measurements used here (see [47] for an in-depth review and [48] for a recent example). Duewer et al. [49] estimated the uncertainty

in factor analysis based on the measurement uncertainty in chemical data. Babamoradi et al. [50] performed an uncertainty analysis of the results obtained by PCA using the bootstrap method. The same authors also used the bootstrap method to calculate confidence limits for control charts in PCA-based batch multivariate statistical process control [51]. Preisner et al. [52] used the bootstrap and jackknife methods to estimate bias and variance for non-supervised and supervised discrimination models for microorganism data. Conlin et al. [53] used a methodology based on bootstrap to estimate the standard deviations of the loading matrix to define confidence bounds for contribution plots using simulated data.

Other studies have investigated PLS-DA to estimate uncertainty or reliability of classified samples. Perez et al. [54, 55] used PLS-DA on publicly available data sets to classify the samples and obtain an expression for the reliability of classification. Botella et al. [56] used microarray data and probabilistic discriminant least squares with reject option to classify biological samples. Almeida et al. [29, 30] used PLS-DA with uncertainty estimation to classify banknotes [30] and Amazonian rosewood essential oil [29]. There are other algorithms that can be used to calculate the measurement uncertainty or reliability reported in literature. Appel et al. [57] in which different algorithms were used to express the probabilistic class identification using metabolomics profiles. We used the work of Almeida to conduct uncertainty analysis on a molecular-structure model of biodegradability developed with PLS-DA [31].

2 Computational methods

2.1 Overview

As discussed in the introduction, this paper presents a meta-analysis of NMR data using a workflow that consists of several previously published components combined into a single pipeline. The workflow begins with a consensus-analysis algorithm that ensures all of the NMR data being analyzed are drawn from the same distribution [28]. The resulting consensus data set is then analyzed using principal components analysis (PCA) with bootstrapping [50] and partial least squares discriminant analysis (PLS-DA) [58] with residual bootstrapping [29, 30]. The overall procedure is shown graphically in Fig. 1, including outlier detection and the uncertainty analysis. The PCA bootstrap is shown graphically in Fig. 2, and the PLS bootstrap in Fig. 3.

2.2 Outlier detection and consensus analysis

The consensus analysis algorithm from [28] was designed to ensure that all measurements are drawn from similar distributions, or in other words that all measurements are of the same fundamental property. In this procedure, the distance between each pair of spectra in a class is measured, here using the Jensen-Shannon divergence d_{JS} [59], given by

$$d_{\mathrm{JS}}\left(\mathbf{x},\mathbf{y}\right)=\sqrt{d_{\mathrm{KL}}\left(\mathbf{x},\mathbf{m}\right)+d_{\mathrm{KL}}\left(\mathbf{y},\mathbf{m}\right)}\ , \tag{1}$$

where **x** and **y** are two spectra, normalized so they integrate to 1, **m** is the arithmetic mean of **x** and **y**, and d_{KL} is the Kullback-Leibler divergence, given by

$$d_{\mathrm{KL}}\left(\mathbf{x},\mathbf{y}\right)=1\Big/\ln 2\sum_{k=1}^{K}x_k\ln\left(\frac{x_k}{y_k}\right)\ . \tag{2}$$

where K is the number of spectral elements and the subscript k denotes a particular spectral element. This results in a matrix of pairwise distances for each class, which is compressed into a vector by taking the row average. The distances are then fit to a lognormal distribution and scored based on that distribution. These scores are concatenated into a matrix **Z**, where Z_{ij} denotes the score of spectrum i measured by laboratory j relative to the other laboratories measuring spectrum i. The statistical distance t_i is defined by taking the Euclidean norm across each column of **Z**,

$$t_i = \sqrt{\sum_{j=1}^{J} Z_{ij}^2} \tag{3}$$

and then this vector **t** is fit to a lognormal distribution and scored, resulting in a laboratory-level score vector **z**. The consensus data set is identified by removing laboratories with z_i values greater than 5.2, corresponding to the 95 % confidence interval for the lognormal distribution.

2.3 Bootstrap uncertainty estimation

Bootstrapping is used to estimate uncertainty in a statistical model in terms of some kind of confidence limit. This approach has been widely documented [38-43, 60]. The bootstrap procedure involves creating a large number of new artificial data sets by randomly choosing members from the true data set with replacement, then fitting the model to these artificial sets. Statistical analysis on these resampled sets can be used to calculate CIs and therefore uncertainties, mainly, in situations where sources of uncertainty are difficult to estimate, such as complex mathematical models, or even in cases where existing sources of uncertainties are not considered in experimental analysis, such as dark uncertainty [61]. In these cases, the bootstrap methodology emerges as an effective means of estimating the uncertainty. Some refinements have been made to the bootstrap methodology, including the bootstrap Latin partition method (BLP) [62], which uses stratification so that every bootstrap sample contains the same proportion of spectra from each group. Such a

method ensures that every group is represented in every bootstrap, although it does not allow an estimate of the uncertainty due to varying the relative sizes of each group.

2.4 Principal components analysis and uncertainty estimation

2.4.1 Principal components analysis

Principal components analysis (PCA) has been one of the most widespread chemometric methods used in chemical sciences [63]. The PCA model can be briefly described through Eq 1,

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{4}$$

where **X** is the array of mean-centered raw spectral data with $I \times K$ matrix, where I is the number of measurements and K the number of spectral variables. **T** is the projection of **X** onto the subspace whose basis vectors form the orthogonal matrix **P**, and **E** is the array of residuals. In PCA, **T** is commonly called the score matrix and **P** the loadings matrix. **T** is $I \times L$, where L is the number of components selected for the decomposition, and **P** is $K \times L$, ordered column-wise by the size of the corresponding eigenvalue. L is the free parameter in the PCA model. If $L = R$, where R is the rank of **X**, then the PCA model is exact and **E** is zero. For $L < R$, then the PCA model becomes inexact and E is no longer identically zero. However, this reduction in dimensionality often results in interpretable results using just a few principal components rather than the full rank data set.

2.4.2 Uncertainty estimates in PCA with bootstrap

When bootstrapping is applied to the PCA, it results in a set of resampled data matrices **X***, each of whose I rows represent a random selection of the rows of **X**, and corresponding resampled scores **T***, resampled loadings **P***, and resampled residuals **E***. Statistical analysis on the set of **T*** and **P*** matrices allows the determination of uncertainty in the model.

The following procedure is used to assign uncertainty to the PCA scores, as adapted from [50]. For each resampled mean-centered data matrix **X***, a PCA model and corresponding **P*** is calculated; the dimensionality of this PCA is the same as the original, meaning that **P** and **P*** have the same dimensions

$$\mathbf{X}^{*} = \mathbf{T}^{*}\mathbf{P}^{*\mathrm{T}} + \mathbf{E}^{*} \tag{5}$$

Then, the new scores matrix $\mathbf{T}_{\mathrm{proj}}$ is determined by projecting **X** into the bootstrap space with

$$\mathbf{T}_{\mathrm{proj}} = \mathbf{X}\mathbf{P}^{*\mathrm{T}}\mathbf{R}^{\mathrm{T}}. \tag{6}$$

PCA spaces are unique only to within reflections, so the transformation **R** is calculated using an orthogonal Procrustes algorithm that aligns $\mathbf{T}_{\mathrm{proj}}$ with **T**. The CIs on **T** are defined by fitting a Hotelling T^2 distribution to the population of $\mathbf{T}_{\mathrm{proj}}$ matrices. The uncertainties in **P** are calculated from calculating CIs for each loading element in the transformed bootstrap, $\mathbf{P}_{\mathrm{proj}} = \mathbf{R}\mathbf{P}^{*}$.

2.5 Partial least squares discriminant analysis and uncertainty estimation

2.5.1 Partial least squares

Partial least squares (PLS) [58] begins by assuming that the independent mean-centered variables **X** and dependent variables **Y** are related by

$$\mathbf{Y} = \mathbf{X}\mathbf{W}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{7}$$

where **Q** and **F** are the loadings and residuals of **Y**, respectively, and **W** is a projection of **X** into a subspace that is a good predictor of **Y**, commonly called the rotation matrix. **W** is $K \times L$ and **Q** is

$1 \times L$. L is the number of latent variables (LV), which are analogous to the principal components in PCA. The model predictions $\mathbf{Y}_{\text{pred}}$ are then given by

$$\mathbf{Y}_{\text{pred}} = \mathbf{XWQ}^{\text{T}} \tag{8}$$

In the case of PLS-DA, the **Y** values are class identifiers and we assign them on values of 0 or 1. The $\mathbf{Y}_{\text{pred}}$ vector consists of real numbers and must be interpreted as a class assignment based on some thresholding scheme. The assigned class can be determined by determining a class decision boundary y_{bound} as discussed in our previous work [31]. Then, the actual class assignment is 1 if the corresponding element of $\mathbf{Y}_{\text{pred}}$ is greater than y_{bound} and 0 otherwise.

2.5.2 Uncertainty estimates in PLS-DA with residual bootstrap

Uncertainty estimation using the PLS-DA model is done using the residual bootstrap method proposed by Almeida [30], as implemented in our previous work [31]. Briefly, the residual bootstrap treats the model residuals as representative of the uncertainty in the model. These residuals are randomly sampled and added back onto the model estimates, thus generating a new set of model values to be estimated. This procedure differs from bootstrap in that the bootstrap sample set contains the same measurements as the original set, just with different **Y** values. The weighted residual $\mathbf{F}_{\text{weight}}$ is calculated as per Almeida [30],

$$\mathbf{F}_{\text{weight}} = \frac{\mathbf{F}}{\sqrt{1 - D_{\text{f}} / I}} \tag{9}$$

where D_f is the number of pseudo-degrees of freedom described by van der Voet [64], given by $D_f = I\sqrt{1-E_{rms}/E_{rmscv}}$, where E_{rms} and E_{rmscv} are the root-mean-squared error of calibration and cross-validation, respectively. A new dependent variable matrix **Y*** is generated by sampling from the weighted residuals and adding those resampled residuals to the model predicted $\mathbf{Y}_{pred}$ values. The residuals are assumed to be representative of the uncertainty in the model, and so a new random residual vector $\mathbf{F}^*$ is generated by bootstrapping the residuals. The $\mathbf{Y}_{pred}$ values are perturbed by adding the bootstrapped residual, so that

$$\mathbf{Y}^* = \mathbf{Y}_{pred} + \mathbf{F}^* . \tag{10}$$

A new PLS-DA model is then calculated which has a new set of scores **Q*** and weights $\mathbf{W}^*$. The bootstrap predicted values $\mathbf{Y}^*_{pred} = \mathbf{X}\mathbf{W}^*\mathbf{Q}^{*T}$ is calculated, along with the difference between $\mathbf{Y}^*_{pred}$ and $\mathbf{Y}_{pred}$, denoted $\hat{\mathbf{F}}$,

$$\hat{\mathbf{F}} = \mathbf{X}\left(\mathbf{W}^*\mathbf{Q}^{*T} - \mathbf{W}\mathbf{Q}^{T}\right) . \tag{11}$$

The CIs on each PLS-DA model prediction are then calculated by determining the percentiles of each element of $\hat{\mathbf{F}}$, denoted $\hat{F}_a$. For the 95 % CI, the lower confidence limit $c_{l,a}$ is the 2.5 percentile and the upper limit $c_{u,a}$ is the 97.5 percentile. We choose percentile as opposed to other methods [42, 65] because it makes no assumptions about the underlying distributions of the data; furthermore, the CIs are often asymmetric, suggesting that the predictions are not normally distributed.

2.5.3 Misclassification probability in PLS-DA

When uncertainty is considered in the case of a two-class discriminant analysis, the result is that there are actually three classes [31]. In addition to the two classes, here labelled as 0 and 1, there is an additional class of "unsure." Samples can be assigned to the "unsure" class based on whether their CIs include y_{bound}. It is this additional "unsure" classification that motivates the misclassification probability. It should be noted that the misclassification probability presented here is not new; it has been presented completely in our previous work [31]. Additional details of the discussion are presented here. The misclassification probability of a sample provides a measure of trustworthiness for the classification of that sample. To calculate this probability, the model-predicted values $\mathbf{Y}_{\text{pred}}$ are treated as normally distributed random variables with mean $\bar{\mathbf{Y}}$ given by the PLS predictions,

$$\bar{\mathbf{Y}} = \mathbf{X}\mathbf{W}^{*}\mathbf{Q}^{*\mathrm{T}}, \tag{8}$$

Each prediction $Y_{a,\text{pred}}$ has a standard deviation σ_a given by

$$\sigma_a = \tfrac{1}{4}\left(c_{\mathrm{u},a} - c_{\mathrm{l},a}\right), \tag{9}$$

where *c* values are the upper and lower 95 % confidence limits from the previous section. The CIs may not be symmetric but they are usually close enough for this to be a reasonable approximation. We then approximate $Y_{a,\text{pred}}$ as being normally distributed, that is, $Y_{a,\text{pred}} \sim N\left(\bar{Y}_a, \sigma_a\right)$. The probability of a sample, indexed by *a*, being assigned to class 0 is then equal to the probability that its predication $Y_{a,\text{pred}}$ is less than the decision threshold y_{bound},

$$\mathrm{Pr}_{0,a} = \mathrm{Pr}\left(Y_{a,\mathrm{pred}} \le y_{\mathrm{bound}}\right) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{y_{\mathrm{bound}} - \bar{Y}_a}{\sqrt{2}\sigma_a}\right)\right] \tag{10}$$

(see [31]). The probability the sample being assigned to class 1 is

$$\mathrm{Pr}_{1,a} = 1 - \mathrm{Pr}_{0,a}\,. \tag{11}$$

If the true class Y_a is known, then the misclassification probability $\mathrm{Pr}_{\mathrm{misclass}}$ is then equal to the probability of the wrong class being chosen, which is

$$\mathrm{Pr}_{\mathrm{misclass},a} = \mathrm{Pr}_{1-Y_a,a}\ . \tag{12}$$

For a new sample with no known true class, Y_a in Eq. 12 can be replaced with the corresponding assigned value. Such samples must then have $\mathrm{Pr}_{\mathrm{misclass}} < 0.5$, and the probabilistic class can only ever be 0, 1, or "unsure"; it can never be "definitely wrong" because there is no other truth to compare with.

The misclassification probabilities can be used to assess the trustworthiness of the model. If the "unsure" classification means that a sample's 95 % CI includes y_{bound}, then the sample is in this "unsure" class if $\mathrm{Pr}_{\mathrm{misclass}} > 0.025$. Additionally, for the samples with a known true class, there is a fourth class, which might be called "definitely wrong". This class would correspond to a sample being misclassified by the model *and also* the 95 % CI *not* including y_{bound}, equivalent to $\mathrm{Pr}_{\mathrm{misclass}} > 0.975$. For a model to be trustworthy, all its misclassified samples would be labelled "unsure" rather than "definitely wrong", and as few as possible correctly-classified samples labelled "unsure". A model with a large number of "definitely wrong" training and test samples would not be trustworthy, because there is a high likelihood that new samples will be classified incorrectly

and be assigned a low $Pr_{misclass}$. Similarly, a model that classifies most samples as "unsure" is also not trustworthy, even if no samples are "definitely wrong", because such a model would never make a prediction that could be relied on.

We present some examples of hypothetical misclassification probabilities in Table 1. In this table, samples A-D are true class 0, samples E-F are true class 1, and samples I-K are of unknown class. Samples A-F are being used to validate the model, and so their classes have been determined by other means. Samples I-K are being assigned to a class based on the results of this model, and so there is no true class assignment available. For samples of known true class, $Pr_{misclass}$ is simply the probability of assignment to the opposite class. For instance, Sample A has a Pr_1 of 0.01. As this is less than 0.025, the sample is probabilistically assigned to class 0. Sample B has a Pr_1 of 0.25, which puts it in the probabilistically unsure class. Sample D has a Pr_1 of 0.98, which is greater than 0.975 and so the sample is probabilistically classified as "definitely wrong". Because Samples I-K have unknown true class, $Pr_{misclass}$ for them is the probability of being assigned to the opposite class from their assigned class. Hence, Sample K is assigned to class 1, but it has a Pr_0 of 0.49 and assigned to the "unsure" class.

Table 1. Example misclassification probabilities and corresponding probabilistic interpretations

Sample	Pr_0	Pr_1	Assigned class	True class	$Pr_{misclass}$	Probabilistic class
A	0.99	0.01	0	0	0.01	0
B	0.75	0.25	0	0	0.25	Unsure
C	0.49	0.51	1	0	0.51	Unsure
D	0.02	0.98	1	0	0.98	Definitely wrong
E	0.02	0.98	1	1	0.02	1
F	0.01	0.99	1	1	0.01	1
G	0.98	0.02	0	1	0.98	Definitely wrong
H	0.99	0.01	0	1	0.99	Definitely wrong
I	0.01	0.99	1	Unknown	0.01	1
J	0.98	0.02	0	Unknown	0.02	0
K	0.49	0.51	1	Unknown	0.49	Unsure

3 Data and Implementation

3.1 Experimental data

Two data sets were used in this study, both taken from the interlaboratory comparison study in Viant et al [66]. In this study, eighteen metabolite mixtures were sent to the participating laboratories, and for each sample, ten one-dimensional ^{1}H NMR spectra were obtained on different instruments across a range of NMR field strengths. In both in the present study and our previous work [28], each instrument has been treated as being completely independent of the others. We know that some laboratories contain more than one instrument, which could introduce a correlation between instruments in the same laboratory. In [28], we determined that the dominant factor influencing the outlier analysis was the NMR field frequency, so the correlation introduced by being in the same laboratory is probably small. The spectra are reported as chemical shift frequencies in parts per million (ppm), with a range from 10.0 ppm to 0.2 ppm, with a region from 4.7 ppm to 5.2 ppm excluded due to water solvent suppression artifacts. The spectra are binned with a bin width of 0.005 ppm, for a total of 1860 variables in each spectrum. Due to additional water suppression artifacts apparent in the spectra, for this analysis, an additional region from 4.2 ppm to 4.7 ppm was excluded, after which the NMR spectra were renormalized such that the sum across each spectrum was 1.

The set of synthetic data from Viant [66] consists of six synthetic metabolite mixtures (S1-S6), each containing the same metabolites in various controlled mixtures; these random concentrations were not designed to form a two-class sample set. One mixture, S1, has six replicates, for a total of 11 samples and 110 spectra. The set of biological data consists of 12 liver extracts from European flounder, six obtained from an unpolluted control site (BC1-BC6) and six from a

polluted site (BE1-BE6), forming, in principle, a two-class sample set. One sample, BC1, has three replicates, for a total of 14 samples and 140 spectra.

3.2 Outlier detection and consensus analysis

In our previous study [28] on this set of data, we performed an outlier-detection analysis to determine if there was a subset of the data that was likely to be more internally consistent than the complete data set. The outliers that we identified were the 800 MHz NMR spectra and one of the 600 MHz spectra showed a shifting of its spectral features, which made it difficult to compare it with the other spectra. The remaining data form the consensus data set. a set of seven laboratories, with 77 synthetic-mixture spectra in total to be compared using PCA and 98 biological-sample spectra in total to be compared using PLS-DA. The identification of the 800 MHz data as an outlier does not mean that the data is necessarily wrong, as we noted in our previous work [28], but rather that it contained information not present in the lower-frequency spectra.

In a statistical sense, this study identified measurements that are likely drawn from a different distribution from the consensus. This is an automated process that does not use any human expertise to curate the measurements. That study found a subset of seven instruments that consistently produced NMR spectra close to consensus. In this paper, therefore, the chemometric analysis is performed only on the spectra from these seven instruments. The code used to perform this analysis is available in the supplemental information of our previous study [28].

3.3 Implementation of PCA and PLS

The PCA and PLS models from scikit-learn 0.18 [67, 68] were used. The synthetic mixtures were analyzed using PCA and the biological extracts were analyzed using PLS-DA. For the synthetic mixtures, a dummy matrix **Y** was created with 0 identifying the control samples and 1

identifying the exposed samples. The number of components for the PCA model was chosen by finding the first *n* components that together explained > 95% of the variance in the NMR spectra. The optimal number of LVs in the PLS-DA model was determined by using leave-one-out cross-validation and minimizing RMSECV. Uncertainty in the PCA model was estimated using bootstrapping (Section 2.4) and that in the PLS model using residual bootstrapping (Section 2.5). We used 1000 bootstrap samples for this study, but we performed additional tests with 10 000 bootstrap samples and determined that the results are not strongly dependent on the number of samples. The procedure used here is similar to that used in our previous paper on PLS-DA uncertainty analysis [31] which has been adapted to the PCA model, and the code is included in the supplementary information.

4 Results

4.1 Synthetic metabolite mixtures and PCA

The bootstrapping methodology was first applied to the PCA of the synthetic mixtures. In the Viant study, the mixtures were chosen so that they would be easily distinguishable merely by examining their NMR spectra, which are shown in Fig. 4. Because the spectra can be visually distinguished, the results of the PCA can be easily tested.

As noted in the introduction, having an idea of the uncertainty in a chemometric analysis is crucial in order to properly understand the results. This motivates the uncertainty analysis on the PCA using bootstrapping, and the results of this analysis are shown in Fig. 5. For this analysis, six PCs are used. These PCs together explain around 97 % of data variance. We knew already that there were six substances in the synthetic samples, which corroborates the choice of how many PCs to retain.

There are two possible estimates of the uncertainty in the PCA predictions, one coming from the scatter in the groups and one coming from the bootstrap analysis. For each group, a Hotelling T^2 [69] 95 % confidence ellipse can be drawn based either on the scatter in the group or the scatter in the bootstrap samples for that group; these are shown in Fig. 5a and 5b. Figure 5a has the 95 % ellipses based on group size only, and Fig. 5b shows the 95 % ellipses based on both group size and bootstrap. The effect of the bootstrap analysis is to introduce a finite size to each sample's location in PCA space, so we would expect the T^2 ellipse to be larger when calculated from bootstrap than from group size, especially when the group size is small relative to the bootstrap scatter size as in groups S1 and S2. However, since the size of the T^2 ellipse is proportional to an *F*-statistic based on the number of samples in each group, the T^2 ellipse based on group size can be larger than that based on bootstrap, especially if the number of samples in a group is small, as in groups S4 and S5. The bootstrap scatter size comes from the large number of bootstrap samples and already contains information about how the PCA is affected by group size; consequently, the *F*-statistic term has less effect.

In Fig. 5a, the 95 % confidence limits for all groups show a statistically significant separation. The S1 and S4 groups lie close to each other, as do the S2 and S6 groups. The proximity of the groups suggests that added uncertainty in the PCA could cause them to merge, and this is indeed what we see in Fig. 5b. The S1 and S4 groups merge due to the additional uncertainty from the bootstrap, and the S2 and S6 almost merge.

Because the uncertainty in the classification will depend strongly on the number of PCs in the PCA model, it is worthwhile to examine what the uncertainties are in the PCA loadings (**P** values from Eq. 4) of each PC, calculated from the bootstrap results (Eq. 6). The loadings and uncertainties for the six PCs are shown in Fig. 6. The first two PCs show strong influence from the

spectral features of nicotinic acid, glucose, citrate, and glutamine. These substances are principally responsible for the differences among the samples. By comparing the uncertainties with the loadings values, we can get an estimate of how significant any particular feature in the loading is. The loading uncertainties for the first three PCs are much smaller than the spectral features (~0.2 for the nicotinic acid peaks as opposed to 0.03 for their uncertainty), which suggests that these PCs contain diagnostic information rather than simply instrument-to-instrument variability. The fourth and fifth PCs have larger uncertainties relative to the feature size, but the features are still much larger than the uncertainties (~0.4 for the alanine peaks as opposed to 0.15 for their uncertainties, for instance). Conversely, the uncertainties in the sixth PC are larger than the spectral features, suggesting this PC is mostly instrument variability or other 'noise'.

4.2 Biological metabolite samples and PLS-DA

Unlike the synthetic mixtures, the biological liver samples are difficult to distinguish visually. A representative sample of the NMR spectra are shown in Fig. 7; the spectral features around 3 ppm to 4 ppm contains much of the information about the samples but is difficult to interpret without chemometric methods. In our previous study, we showed that PCA could shed some light on the differences between the classes. Here, we show that PLS-DA can be used to separate the samples based on whether the fish were from the control or exposed sites and to identify the spectral features responsible for that separation. As with PCA, bootstrapping was used to calculate the uncertainties in the classifications, scores, and LV loadings; in this case, the residual bootstrap method was used as discussed in Section 2.4. The number of LVs was determined by finding the first minimum in the root mean squared error of cross-validation, which yielded two LVs. These two variables together explain 90 % of the variance among the samples.

The most significant output of the model is how it predicts the class of the liver samples, which is shown in Fig. 8. This figure shows both the predicted classes and the uncertainty in those predictions. Every sample is classified correctly, although the 95 % CIs sometimes reach close to y_{bound} so that $Pr_{misclass}$ is finite but still relatively small. This model has a Pearson's R^2 value of 0.87 and a root mean squared error of 0.177 for the predicted Y values. The result can be compared to fitting a PLS model to the full data set without using the consensus analysis procedure; this model is shown in Fig. 9. Two samples are classified incorrectly, one of which is definitely misclassified. Three additional samples, although classified correctly, are "unsure." The R^2 value is 0.80 and the RMS error is 0.222. These results are not terribly worse than the results using the consensus analysis procedure. Since there is no test set, however, it cannot be proven that the models are not overfitting the data. It is this concern which motivates the cross-validation test.

In cross-validation, some laboratories are held out as a test set while the PLS-DA model is fit against the remaining data, the training set. The accuracy and precision of the model is then judged based on how well it performs on the test set. In this case, the test set consisted of three randomly-selected laboratories and the training set of the remaining four. The PLS-DA classifications for the test and training sets are shown in Fig. 10, along with $Pr_{misclass}$ values; likewise, we perform the same test using all labs, without using the consensus analysis procedure, and the results are shown in Fig. 11. When using consensus analysis, for both the test set and training set, one sample is misclassified, and some other correctly-classified samples have error bars large enough to include y_{bound}. As such, these samples have a $Pr_{misclass}$ close to 0.5, indicating that the model assigns a low degree of assurance to these classifications. The R^2 value for classification is 0.82, which is also not much worse than without the cross-validation procedure. As an additional test, we repeated the cross-validation ten times, the results of which are shown in supplementary Fig. S1. Never more

than three samples are misclassified in these validation tests, and none ever definitely wrong. Almost all samples are assigned low $Pr_{misclass}$. We argued in Section 2.8 and in our previous PLS-DA study [31] that a model could be considered reliable if it assigned low misclassification probability to correctly-identified samples and $Pr_{misclass}$ close to 0.5 for misidentified samples. The model here does exactly that, and so the model can be said with some assurance to be reliable.

The situation is much different when using the complete data set, without consensus analysis. The R^2 value is 0.69, considerably worse than using consensus analysis. Furthermore, nine samples from the training and test sets are misclassified, including three as definitely wrong. We repeat the test ten times, showing the results in Fig. S2; the results are similar, with never fewer than three misclassified samples and, in almost every case, at least one definitely wrong. Therefore, the model trained on the complete data set cannot be said to be reliable.

In addition to how the model classifies the liver samples, it is important to understand the physical basis assigned by the model for those classifications. This information comes from interpreting the latent variable loading values (the pseudoinverse of the **W** values from Eq. 4), which are shown in Fig. 12. This figure also shows the uncertainties in the latent variables, which can help to explain the amount of importance that should be assigned to any particular feature. For the first LV (78% explained variance (EV) in **X**), the uncertainties in the loadings are almost zero compared to the loadings themselves (~0.2 for the loadings as opposed to ~0.01 for the uncertainties). For the second LV (22% EV in **X**), the relative uncertainties are much larger (~0.2 versus ~0.1); similar to the PCA, the large relative uncertainty suggests that the second LV contains less useful diagnostic information than the first.

In the original Viant study [66], PCA was used to identify several metabolites as being responsible for separating the exposed-site fish from the control-site fish, namely glucose, lactate,

and three unknown substances. Our recent study [28] using PCA combined with outlier analysis identified another spectral feature at a frequency shift of approximately 1.5 ppm, although that feature was not assigned a chemical identity. This same feature appears prominently in the loadings of the first LV, in Fig. 12, indicating that changes in the associated substance can be used to separate the exposed-site and control-site fish. The low uncertainty assigned to this spectral feature suggests that it has some diagnostic value in distinguishing the two groups. Likewise, a similar feature appears in the second LV, but the uncertainty is quite large and so it is likely does not contribute to separation along this axis.

5 Conclusions

In this work, we have proposed a start-to-finish consistency analysis, consensus analysis, and uncertainty analysis workflow for metabolomics and chemometrics more generally. Here, we apply this workflow to separation and classification of environmental metabolomics NMR spectra, but the workflow is general and will be useful for interlaboratory intercomparison analysis. The consistency and consensus analysis portions of the workflow were conducted in an earlier study [28] and the results of that analysis were used as an input to an uncertainty algorithm developed for partial least squares [31].

We used principal components analysis (PCA) and partial least squares discriminant analysis (PLS-DA) to separate and classify environmental metabolomics NMR spectra. The PCA was conducted on samples of specified composition and the PLS-DA was conducted on fish liver samples from an industrially contaminated site and a control site. In all cases, the uncertainty analysis results in a differing interpretation from the situation without uncertainty analysis. When uncertainty analysis was added to PCA, groups of samples that would have been considered as

separate without uncertainty analysis became merged once the uncertainty in the scores was included.

Likewise, the PLS-DA model by itself would have been said to be acceptable without uncertainty analysis. With this uncertainty protocol, it becomes possible to classify samples as exposed-site and control-site, but also to attach a level of assurance to that classification. In particular, even though the PLS-DA model appears to be adequate, the model is able to demonstrate different levels of confidence for some classifications, providing a warning that the classification may be wrong. The model never makes a confident classification that turns out to be wrong ($Pr_{misclass} \approx 1$).

One advantage of the uncertainty analysis performed here is that the bootstrap-based models can be provided in the metadata for the analysis, meaning that anyone can take an unknown sample and use the model to generate a prediction and uncertainty. Consequently, if the model is used to make a prediction in the field, an uncertainty will be assigned to that prediction. The assigned uncertainty means that the prediction can be used with substantially more confidence and also easily compared with an experimental measurement, if independent verification should be necessary. Wider adoption of an uncertainty analysis strategy for model development will help address many challenges faced by chemometric model development.

Disclaimer

Certain commercial equipment, instruments, or materials are identified in this paper in order to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology or the National Institute of Metrology, Quality and Technology, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose.

Acknowledgements

This work was partially supported by the Conselho Nacional de Desenvolvimento Científico e Tecnológico (National Council for Scientific and Technological Development) of Brazil [grant number REF.203264/2014-26]. The authors also thank Dr. Katrice Lippa, Dr. David Duewer, and Dr. Pamela Chu at NIST for productive discussions.

Compliance with Ethical Standards

The authors declare that they have no conflict of interest.

References

1. Nicholson JK, Wilson ID. Understanding 'Global' Systems Biology: Metabonomics and the Continuum of Metabolism. Nat Rev Drug Discov. 2003;2(8):668-76.
2. Lu X, Zhao X, Bai C, Zhao C, Lu G, Xu G. LC–MS-based metabonomics analysis. J Chromatogr B. 2008;866(1–2):64-76.
3. Willenberg I, Ostermann AI, Schebb NH. Targeted metabolomics of the arachidonic acid cascade: current state and challenges of LC–MS analysis of oxylipins. Anal Bioanal Chem. 2015;407(10):2675-83.
4. Karaman İ, Nørskov NP, Yde CC, Hedemann MS, Bach Knudsen KE, Kohler A. Sparse multi-block PLSR for biomarker discovery when integrating data from LC–MS and NMR metabolomics. Metabolomics. 2015;11(2):367-79.
5. Hsu C-C, ElNaggar MS, Peng Y, Fang J, Sanchez LM, Mascuch SJ, et al. Real-Time Metabolomics on Living Microorganisms Using Ambient Electrospray Ionization Flow-Probe. Anal Chem. 2013;85(15):7014-8.
6. Rath CM, Yang JY, Alexandrov T, Dorrestein PC. Data-Independent Microbial Metabolomics with Ambient Ionization Mass Spectrometry. J Am Soc Mass Spectrom. 2013;24(8):1167-76.
7. Weston DJ. Ambient ionization mass spectrometry: current understanding of mechanistic theory; analytical performance and application areas. Analyst. 2010;135(4):661-8.
8. Evans AM, DeHaven CD, Barrett T, Mitchell M, Milgram E. Integrated, Nontargeted Ultrahigh Performance Liquid Chromatography/Electrospray Ionization Tandem Mass Spectrometry Platform for the Identification and Relative Quantification of the Small-Molecule Complement of Biological Systems. Anal Chem. 2009;81(16):6656-67.
9. Ehrhardt C, Arapitsas P, Stefanini M, Flick G, Mattivi F. Analysis of the phenolic composition of fungus-resistant grape varieties cultivated in Italy and Germany using UHPLC-MS/MS. J Mass Spectrom. 2014;49(9):860-9.
10. Rodriguez-Aller M, Gurny R, Veuthey J-L, Guillarme D. Coupling ultra high-pressure liquid chromatography with mass spectrometry: Constraints and possible applications. J Chromatogr A. 2013;1292:2-18.
11. Wishart DS. Quantitative metabolomics using NMR. TrAC, Trends Anal Chem. 2008;27(3):228-37.

12. Viant MR, Lyeth BG, Miller MG, Berman RF. An NMR metabolomic investigation of early metabolic disturbances following traumatic brain injury in a mammalian model. NMR Biomed. 2005;18(8):507-16.
13. Arana VA, Medina J, Alarcon R, Moreno E, Heintz L, Schäfer H, et al. Coffee's country of origin determined by NMR: The Colombian case. Food Chem. 2015;175:500-6.
14. Noothalapati H, Shigeto S. Exploring Metabolic Pathways in Vivo by a Combined Approach of Mixed Stable Isotope-Labeled Raman Microspectroscopy and Multivariate Curve Resolution Analysis. Anal Chem. 2014;86(15):7828-34.
15. Hosokawa M, Ando M, Mukai S, Osada K, Yoshino T, Hamaguchi H-o, et al. In Vivo Live Cell Imaging for the Quantitative Monitoring of Lipids by Using Raman Microspectroscopy. Anal Chem. 2014;86(16):8224-30.
16. Gilany K, Moazeni-Pourasil RS, Jafarzadeh N, Savadi-Shiraz E. Metabolomics fingerprinting of the human seminal plasma of asthenozoospermic patients. Mol Reprod Dev. 2014;81(1):84-6.
17. Dettmer K, Aronov PA, Hammock BD. Mass spectrometry-based metabolomics. Mass Spectrom Rev. 2007;26(1):51-78.
18. Fonville JM, Richards SE, Barton RH, Boulange CL, Ebbels TMD, Nicholson JK, et al. The evolution of partial least squares models and related chemometric approaches in metabonomics and metabolic phenotyping. J Chemom. 2010;24(11-12):636-49.
19. Gromski PS, Xu Y, Correa E, Ellis DI, Turner ML, Goodacre R. A comparative investigation of modern feature selection and classification approaches for the analysis of mass spectrometry data. Anal Chim Acta. 2014;829:1-8.
20. Ouyang M, Zhang Z, Chen C, Liu X, Liang Y. Application of sparse linear discriminant analysis for metabolomics data. Anal Methods. 2014;6(22):9037-44.
21. Wu X, Zhao L, Peng H, She Y, Feng Y. Search for Potential Biomarkers by UPLC/Q-TOF–MS Analysis of Dynamic Changes of Glycerophospholipid Constituents of RAW264.7 Cells Treated With NSAID. Chromatographia. 2015;78(3):211-20.
22. Li Y-Q, Liu Y-F, Song D-D, Zhou Y-P, Wang L, Xu S, et al. Particle swarm optimization-based protocol for partial least-squares discriminant analysis: Application to 1H nuclear magnetic resonance analysis of lung cancer metabonomics. Chemom Intell Lab Syst. 2014;135:192-200.
23. Uarrota VG, Moresco R, Coelho B, Nunes EdC, Peruch LAM, Neubert EdO, et al. Metabolomics combined with chemometric tools (PCA, HCA, PLS-DA and SVM) for screening cassava (Manihot esculenta Crantz) roots during postharvest physiological deterioration. Food Chem. 2014;161:67-78.
24. Heinemann J, Mazurie A, Tokmina-Lukaszewska M, Beilman GJ, Bothner B. Application of support vector machines to metabolomics experiments with limited replicates. Metabolomics. 2014;10(6):0.
25. Wang X, Zhang M, Ma J, Zhang Y, Hong G, Sun F, et al. Metabolic Changes in Paraquat Poisoned Patients and Support Vector Machine Model of Discrimination. Biol Pharm Bull. 2015;38(3):470-5.
26. Tsugawa H, Tsujimoto Y, Arita M, Bamba T, Fukusaki E. GC/MS based metabolomics: development of a data mining system for metabolite identification by using soft independent modeling of class analogy (SIMCA). BMC Bioinformatics. 2011;12(1):131.
27. Dunn WB, Broadhurst DI, Edison A, Guillou C, Viant MR, Bearden DW, et al. Quality assurance and quality control processes: summary of a metabolomics community questionnaire. Metabolomics. 2017;13(5):50.
28. Sheen DA, Rocha WFC, Lippa KA, Bearden DW. A scoring metric for multivariate data for reproducibility analysis using chemometric methods. Chemom Intell Lab Syst. 2017;162:10-20.
29. Almeida MR, Fidelis CHV, Barata LES, Poppi RJ. Classification of Amazonian rosewood essential oil by Raman spectroscopy and PLS-DA with reliability estimation. Talanta. 2013;117:305-11.

30. de Almeida MR, Correa DN, Rocha WFC, Scafi FJO, Poppi RJ. Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation. Microchem J. 2013;109:170-7.
31. Rocha WFC, Sheen DA. Classification of biodegradable materials using QSAR modelling with uncertainty estimation. SAR QSAR Environ Res. 2016:1-13.
32. Gallo V, Intini N, Mastrorilli P, Latronico M, Scapicchio P, Triggiani M, et al. Performance Assessment in Fingerprinting and Multi Component Quantitative NMR Analyses. Anal Chem. 2015;87(13):6709-17.
33. Bich W. Error, uncertainty and probability. In: Bava E, Kuhne M, Rossi AM, editors. Metrology and Physical Constants. 1852013. p. 47-73.
34. Faber K, Kowalski BR. Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler. Chemom Intell Lab Syst. 1996;34(2):283-92.
35. Faber NM, Song XH, Hopke PK. Sample-specific standard error of prediction for partial least squares regression. TrAC, Trends Anal Chem. 2003;22(5):330-4.
36. Fernández Pierna JA, Jin L, Wahl F, Faber NM, Massart DL. Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error. Chemom Intell Lab Syst. 2003;65(2):281-91.
37. Datta J, Ghosh JK. Bootstrap—An exploration. Statistical Methodology. 2014;20:63-72.
38. Kreiss J-P, Paparoditis E. Bootstrap methods for dependent data: A review. J Korean Stat Soc. 2011;40(4):357-78.
39. Wehrens R, Putter H, Buydens LMC. The bootstrap: a tutorial. Chemom Intell Lab Syst. 2000;54(1):35-52.
40. Harrington PB, Laurent C, Levinson DF, Levitt P, Markey SP. Bootstrap classification and point-based feature selection from age-staged mouse cerebellum tissues of matrix assisted laser desorption/ionization mass spectra using a fuzzy rule-building expert system. Anal Chim Acta. 2007;599(2):219-31.
41. Kijewski T, Kareem A. On the reliability of a class of system identification techniques: insights from bootstrap theory. Struct Saf. 2002;24(2–4):261-80.
42. Efron B, Tibshirani RJ. An Introduction to the Bootstrap. New York: Chapman & Hall; 1993.
43. Hjorth JSU. Computer Intensive Statistical Methods: Validation, Model Selection, and Bootstrap. New York: Chapman and Hall; 1993.
44. Olivieri AC, Faber NM, Ferré J, Boqué R, Kalivas JH, Mark H. Uncertainty estimation and figures of merit for multivariate calibration. Pure Appl Chem. 2006;78(3):633-61.
45. Faber K, Kowalski BR. Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. J Chemom. 1997;11(3):181-238.
46. Martens H, Martens M. Modified Jack-knife estimation of parameter uncertainty in bilinear modelling by partial least squares regression (PLSR). Food Qual Prefer. 2000;11(1–2):5-16.
47. Wentzell PD. The Errors of My Ways: Maximum Likelihood PCA Seventeen Years after Bruce. 40 Years of Chemometrics – From Bruce Kowalski to the Future. ACS Sym Ser. 1199: American Chemical Society; 2015. p. 31-64.
48. Karakach TK, Wentzell PD, Walter JA. Characterization of the measurement error structure in 1D 1H NMR data for metabolomics studies. Anal Chim Acta. 2009;636(2):163-74.
49. Duewer DL, Kowalski BR, Fasching JL. Improving the reliability of factor analysis of chemical data by utilizing the measured analytical uncertainty. Anal Chem. 1976;48(13):2002-10.
50. Babamoradi H, van den Berg F, Rinnan Å. Bootstrap based confidence limits in principal component analysis — A case study. Chemom Intell Lab Syst. 2013;120:97-105.
51. Babamoradi H, van den Berg F, Rinnan Å. Comparison of bootstrap and asymptotic confidence limits for control charts in batch MSPC strategies. Chemom Intell Lab Syst. 2013;127:102-11.
52. Preisner O, Lopes JA, Menezes JC. Uncertainty assessment in FT-IR spectroscopy based bacteria classification models. Chemom Intell Lab Syst. 2008;94(1):33-42.

53. Conlin AK, Martin EB, Morris AJ. Confidence limits for contribution plots. J Chemom. 2000;14(5-6):725-36.
54. Pérez NF, Ferré J, Boqué R. Calculation of the reliability of classification in discriminant partial least-squares binary classification. Chemom Intell Lab Syst. 2009;95(2):122-8.
55. Pérez NF, Ferré J, Boqué R. Multi-class classification with probabilistic discriminant partial least squares (p-DPLS). Anal Chim Acta. 2010;664(1):27-33.
56. Botella C, Ferré J, Boqué R. Classification from microarray data using probabilistic discriminant partial least squares with reject option. Talanta. 2009;80(1):321-8.
57. Appel IJ, Gronwald W, Spang R. Estimating classification probabilities in high-dimensional diagnostic studies. Bioinformatics. 2011;27(18):2563-70.
58. Wold S, Sjöström M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemom Intell Lab Syst. 2001;58(2):109-30.
59. Lin J. Divergence measures based on the Shannon entropy. IEEE T Inform Theory. 1991;37(1):145-51.
60. Harrington PdB. Multiple Versus Single Set Validation of Multivariate Models to Avoid Mistakes. Crit Rev Anal Chem. 2018;48(1):33-46.
61. Thompson M, Ellison SLR. Dark uncertainty. Accredit Qual Assur. 2011;16(10):483-7.
62. Wan C, de B. Harrington P. Screening GC-MS data for carbamate pesticides with temperature-constrained–cascade correlation neural networks. Anal Chim Acta. 2000;408(1):1-12.
63. Cardoso Galhardo CE, Rocha WFdC. Exploratory analysis of biodiesel/diesel blends by Kohonen neural networks and infrared spectroscopy. Anal Methods. 2015;7(8):3512-20.
64. van der Voet H. Pseudo-degrees of freedom for complex predictive models: the example of partial least squares. J Chemom. 1999;13(3-4):195-208.
65. Davison AC, Hinkley DV. Bootstrap Methods and their Application. Cambridge: Cambridge University Press; 1997.
66. Viant MR, Bearden DW, Bundy JG, Burton IW, Collette TW, Ekman DR, et al. International NMR-Based Environmental Metabolomics Intercomparison Exercise. Environ Sci Technol. 2009;43(1):219-25.
67. Engel MA. Multiple Objective Resource Allocation in Product and Process Development. Cambridge, MA: Massachusetts Institute of Technology; 1999.
68. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825-30.
69. Massart DL, Vandeginste BGM. Handbook of Chemometrics and Qualimetrics: Elsevier; 1998.

Figures

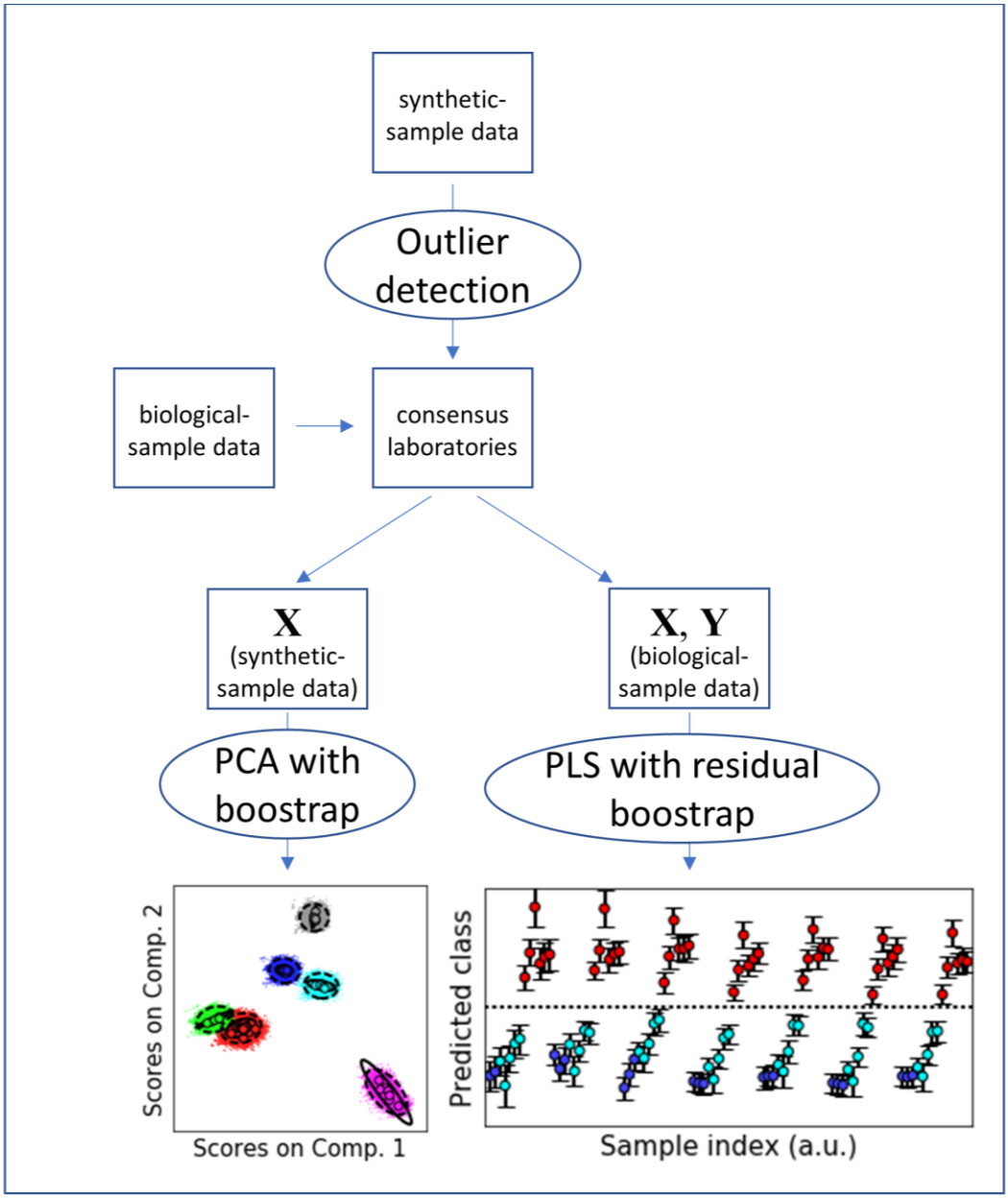


Figure 1. Graphical representation of the overall workflow presented in this paper. Synthetic-sample data is used to determine the consensus set of laboratories. The synthetic-sample data from these laboratories is passed to PCA with bootstrapping, while the biological-sample data is passed to PLS with residual bootstrapping.

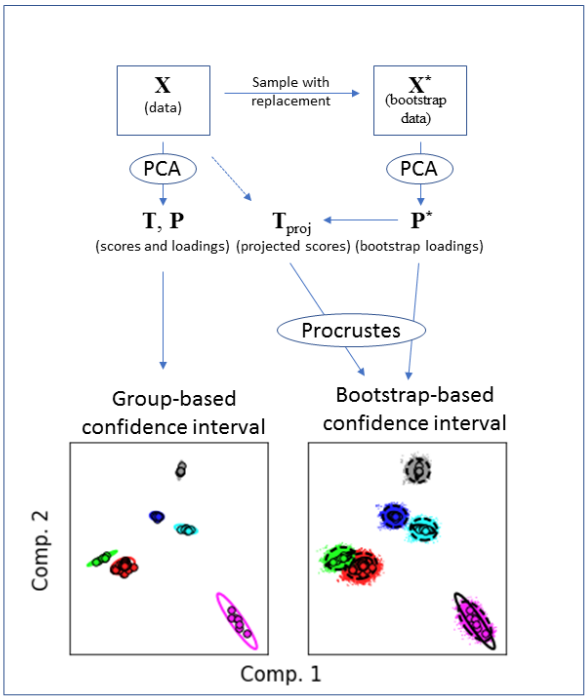


Figure 2. Graphical representation of the PCA bootstrapping process, adapted from [50]. The X data is sampled with replacement, and PCA is performed on each sample. The X data is then projected into the bootstrap PCA space and aligned with the original bootstrap scores. Uncertainty in the scores is estimated based on Hotelling distance within each class.

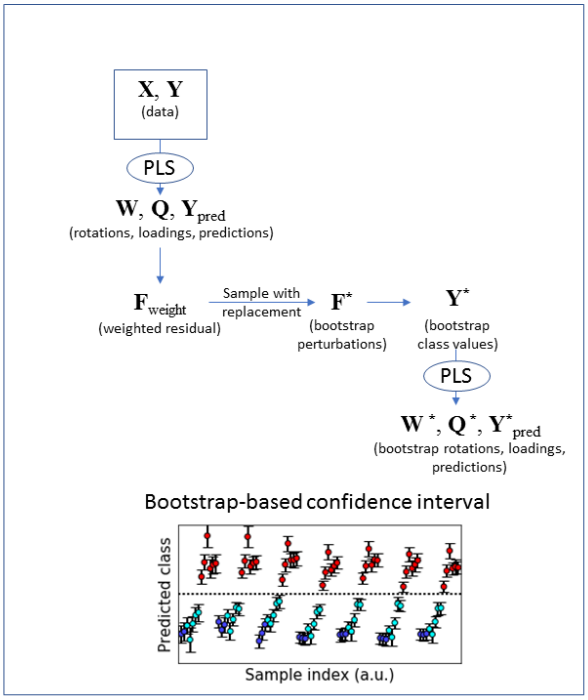


Figure 3. Graphical representation of the PLS residual bootstrapping process. The PLS model is calculated and then a new Y data set is constructed by sampling from the residuals of the original model, after which new models are calculated. Uncertainty in the PLS predictions is estimated from the predictions of the models on each spectrum.

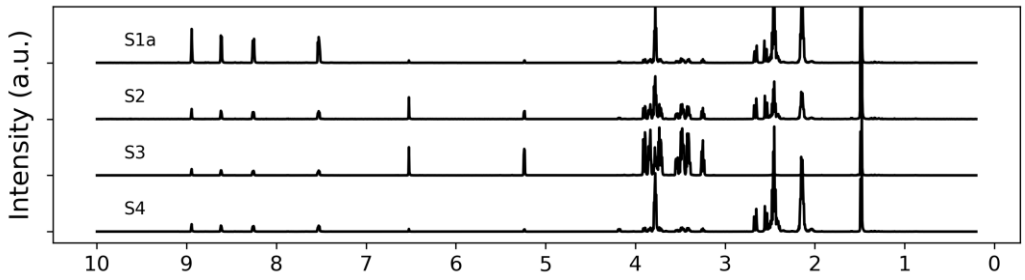


Figure 4. Representative synthetic-mixture NMR spectra with sample labels from Viant, et al. [66]

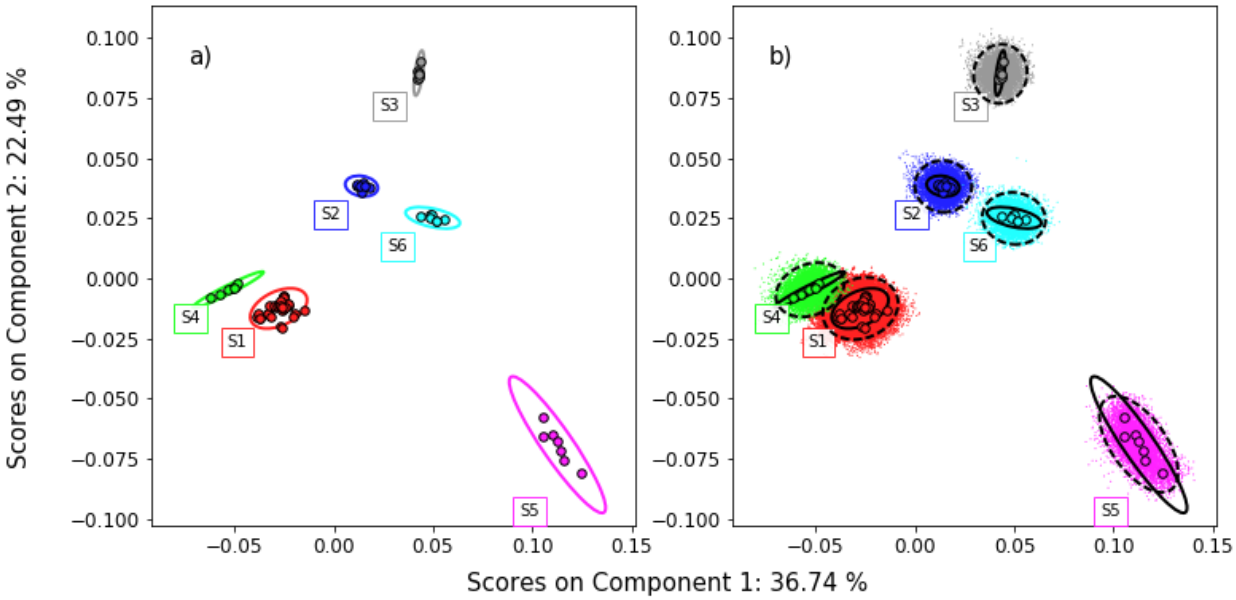


Figure 5. Principal component scores and uncertainties for the synthetic mixture NMR spectra. Uncertainties in the groups are calculated using a the Hotelling T^2 95 % confidence ellipse based on a) the scatter in each individual group (solid ellipse) and b) the scatter in the bootstrap samples (dashed ellipse).

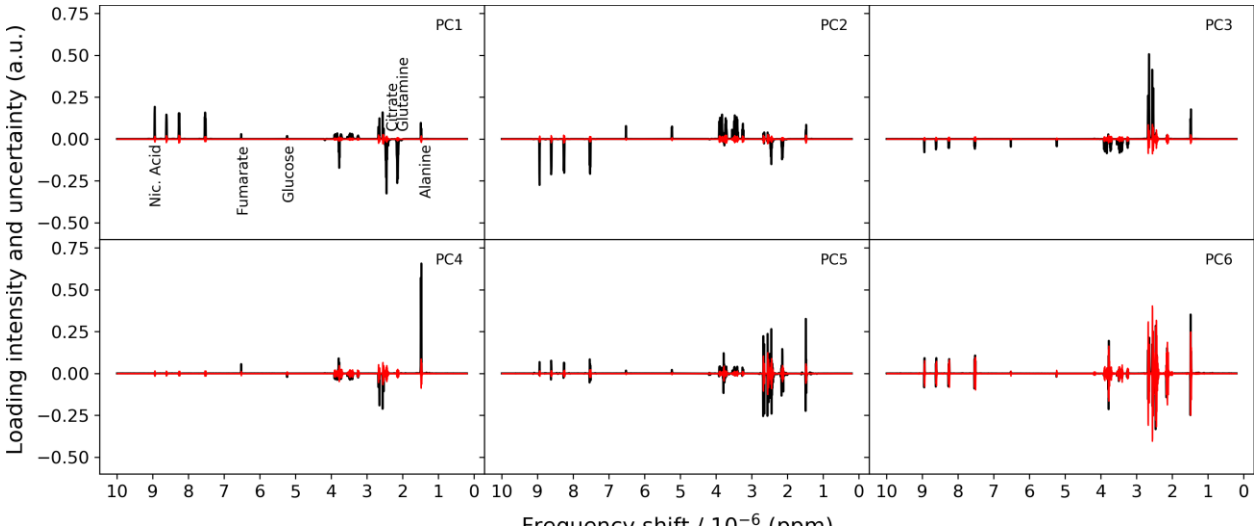


Figure 6. Principal component loadings (black) and uncertainties based on 95 % confidence intervals (red) for the six PCs in the PCA model.

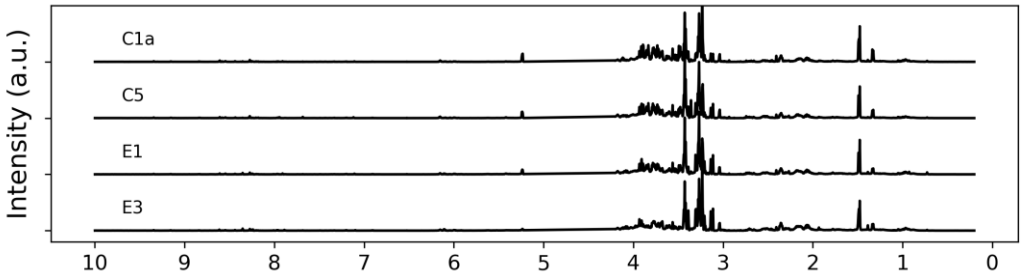


Figure 7. Representative biological-sample NMR spectra with sample labels from Viant, et al. [66].

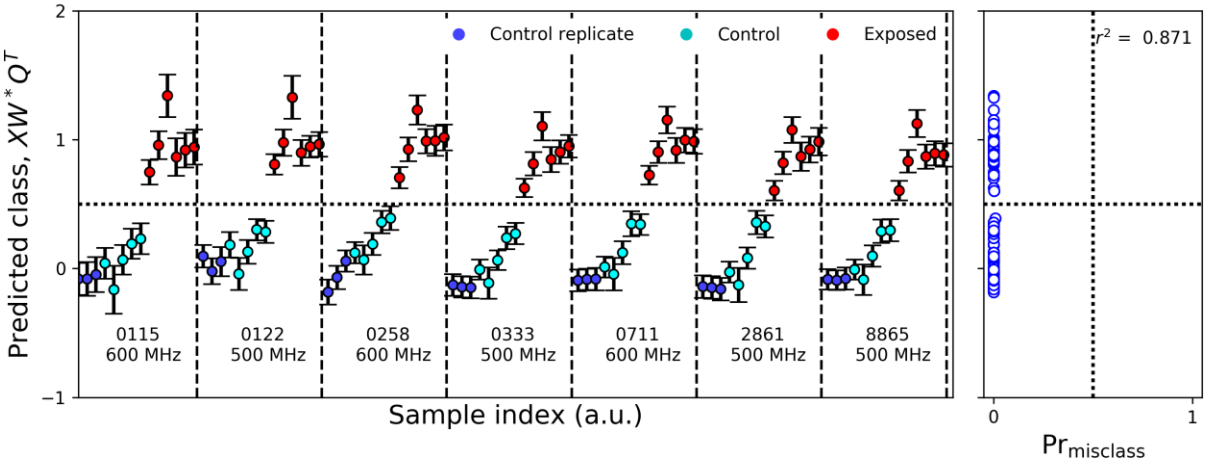


Figure 8. Sample classification (left) and misclassification probability (right) from partial least squares for the biological-sample NMR spectra with 95 % confidence uncertainties calculated from the residual bootstrap. Different laboratories are separated on the plot by vertical dashed lines and are identified by the laboratory number and NMR field strength from Viant, et al. [66]. Data have been screened by the consensus-analysis algorithm [28].

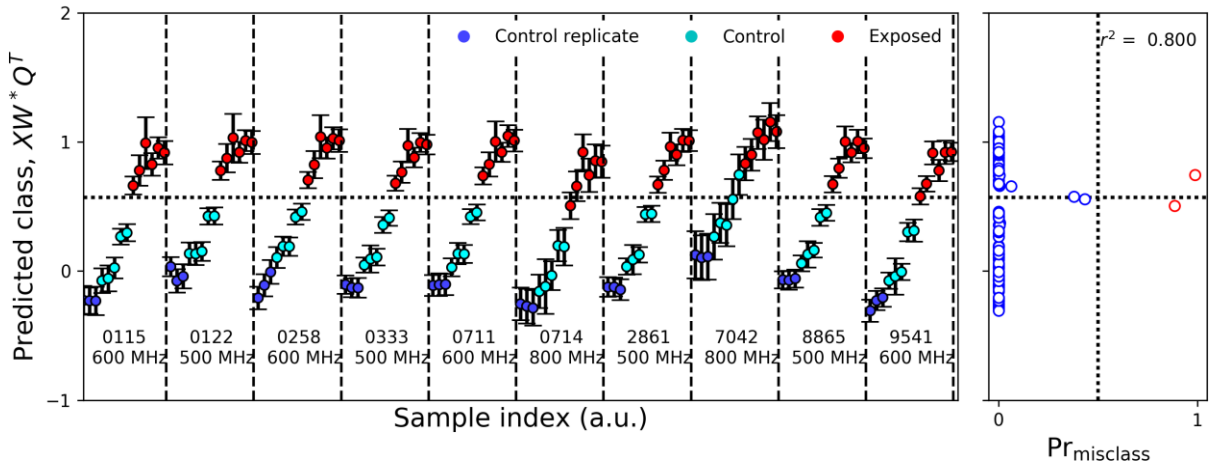


Figure 9. Sample classification (left) and misclassification probability (right) from partial least squares without using the consensus-analysis algorithm [28]. This figure is otherwise identical to Fig. 8.

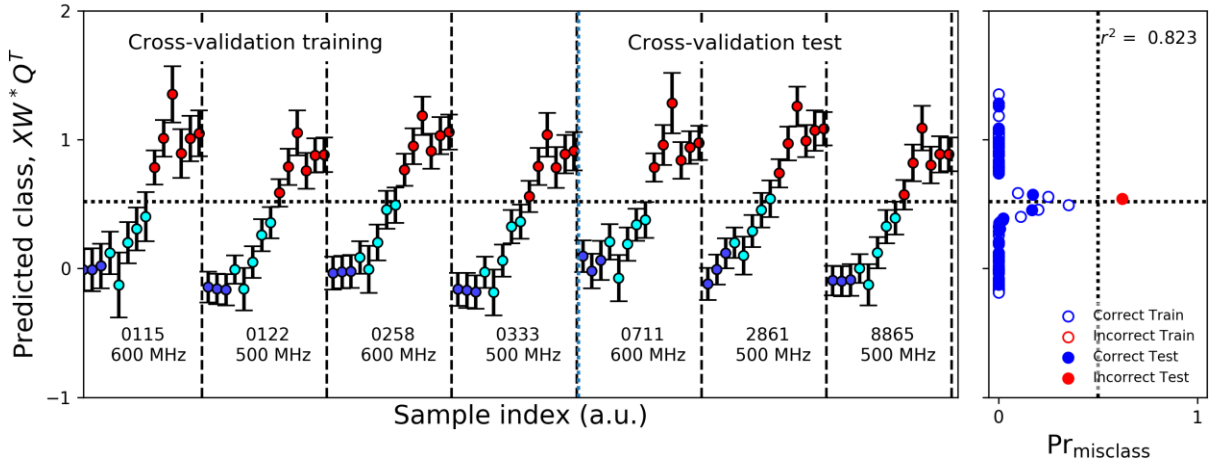


Figure 10. Representative sample classification (left) and misclassification probability (right) from PLS with cross-validation for the biological-sample NMR spectra with 95 % confidence uncertainties calculated from the residual bootstrap. Different laboratories are separated by dashed vertical lines and are identified by the laboratory number and NMR field strength from Viant, et al. [66]. The test and training set are separated by the vertical dash-dot line (between lab 0333 and 0711). Misclassification probabilities for the training set are shown with open circles and for the test set with closed circles. Data have been screened by the consensus-analysis algorithm [28].

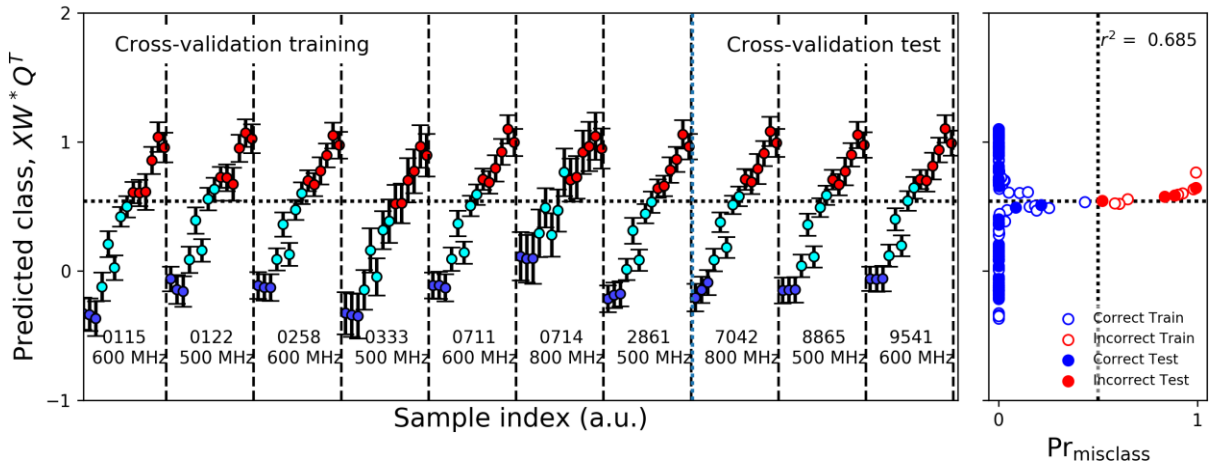


Figure 11. Representative sample classification (left) and misclassification probability (right) from PLS with cross-validation without using the consensus-analysis algorithm [28]. This figure is otherwise identical to Fig. 10

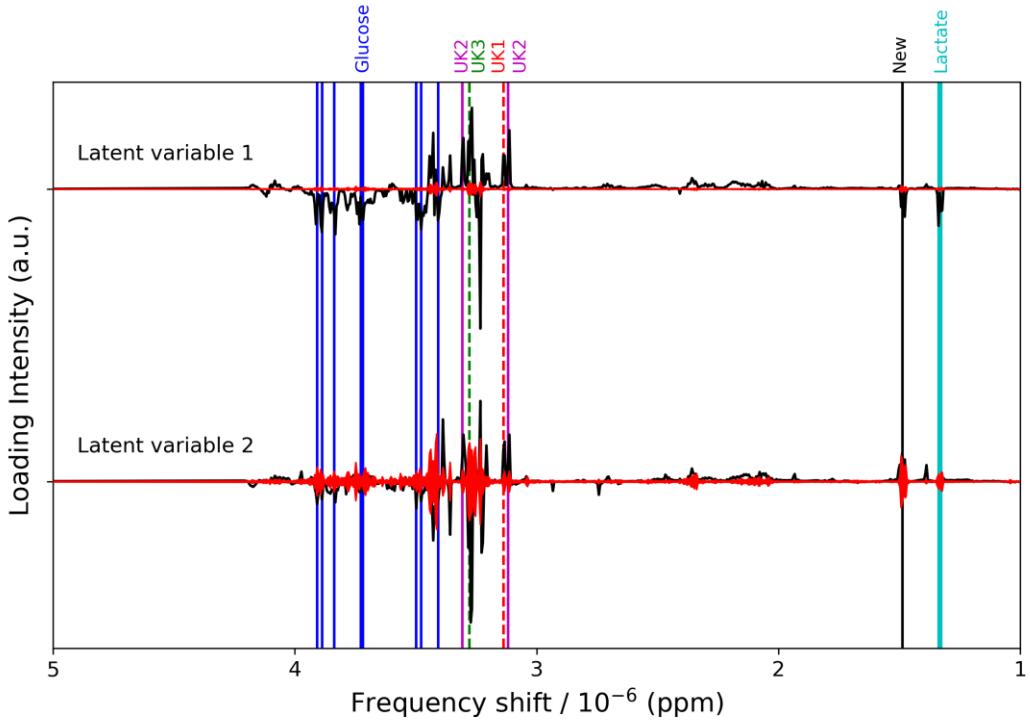
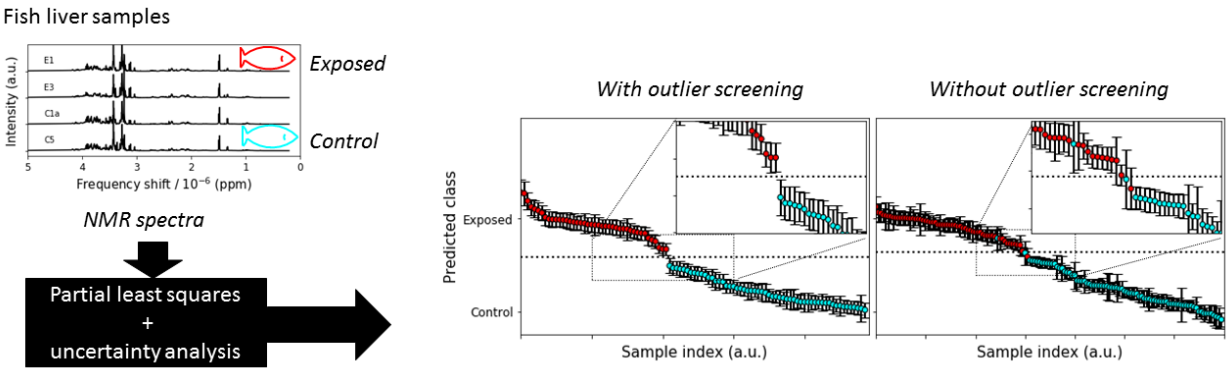


Figure 12. Latent variable loadings (black) and uncertainties based on 95 % confidence intervals (red) for the two LVs in the PLS model. Spectral features identified in the Viant, et al. study are marked with vertical lines. Glucose is blue, lactate is cyan, and the three unknown metabolites are red, magenta (dash-dot), and green (dash-dot).



Graphical Abstract.