Impact of data reduction on multivariate classification models built on spectral data from bio-samples

Journal:	*Journal of Analytical Atomic Spectrometry*
Manuscript ID:	JA-ART-12-2014-000467.R1
Article Type:	Paper
Date Submitted by the Author:	03-Feb-2015

Complete List of Authors:	Larsson, Anders; Swedish Defence Research Agency, FOI, CBRN Defence and Security Andersson, Henrik; Swedish Defence Research Agency, FOI, CBRN Defence and Security Landström, Lars; Swedish Defence Research Agency, FOI, CBRN Defence and Security

Impact of data reduction on multivariate classification models built on spectral data from bio-samples

Anders Larsson, Henrik Andersson, and Lars Landström*

Department of CBRN Defence and Security, SE-901 82 Umeå,

Swedish Defence Research Agency (FOI), Sweden

Abstract

Multivariate data analysis methods have been used to evaluate single shot spectral data, obtained by laser induced breakdown spectroscopy (LIBS), from ten different biological samples (simulants and possible interferents in Biological Warfare Agent (BWA) detection applications). Spectral data as echellograms (2D CCD images) and extracted 1D spectra were used and the classification performance was studied as the number of input variables was altered. Principal component analysis (PCA) indicated a possibility to separate the samples due to spectral differences, and partial least squares discriminant analysis (PLS-DA) was applied to study the predictability in more detail. For full resolution 1D spectra, a normalization of the data mainly resulted in visual effects in the PCA score-plots without significant effect in predictability by the PLS-DA models, however, normalization improved the predictability if the amount of variables were heavily reduced. A quite strong data (variable) reduction could be performed on both the 1D and 2D data without losing significant predictability. Using similar amounts of variables, the prediction models performed better using the echellograms directly compared to the extracted 1D spectra. The problem of spectral data shift (relative 'database' spectra) was also investigated, where already small shifts cause the models to fail. However, after a selection of important variables and allowing certain regions for these variables, the impact of shift on predictability could be reduced.

I. INTRODUCTION

Classification methods and models have become increasingly important in various kinds of, e.g., medical, bioinformatical, detection, etc., applications[1–5]. In our case, it concerns fast and reliable biodetectors based on optical techniques, for "detect-to-warn" purposes in biodefence applications. As there may only be subtle differences in pathogenic and harmless strains of bacteria, combined with commonly relatively low specificity obtained by optical spectroscopy techniques (for biodefence applications), it is of utmost importance to fully evaluate and use the experimental data extracted as well as optimize the experimental set ups to the specific challenge. However, even if some specificity is lost in rapid "detect to warn" set ups such systems may, for example, be useful to initiate a sampling procedure followed by a more time consuming identification step, e.g., Polymerase Chain Reaction (PCR) analysis or other biochemical reaction steps.

Different techniques based on optical spectroscopy are promising as fast detect-to-warn systems and are being continuously improved by extensive research and development efforts (and commercial bioaerosol detection instruments already exist)[6]. Examples of commonly used techniques in biodetection are, e.g., Laser Induced Fluorescence (LIF)[7–9], Flame Emission Spectroscopy (FES)[10], Raman spectroscopy[11,12], and Spark or Laser Induced Breakdown Spectroscopy (SIBS/LIBS)[9,13–15], to mention a few. However, to further reduce the false positive/negative rates in alarm algorithms based on data acquired by the optical techniques mentioned above, some gain can be expected by optimizing which variables to monitor. Such optimization will in many cases increase the sampling rate (to ensure that all samples of interest are analyzed) and also result in better warning algorithms.

In this study, LIBS was used as spectroscopic technique to acquire spectral data from ten different biological samples representing simulants for, e.g., anthrax bacteria, but also a few common background interferent bio-materials such as pollen. Several reports have been published on the use of LIBS for classification on a variety sample sets, e.g., explosives[16], bio-material[14,15,17], chemical warfare simulants/agents[18], and minerals[19]. However, the motivation of the work presented here was mainly to investigate the effect of reducing the number and/or the resolution of sampled data points in the acquired data when performing classification on the sample data set. In addition, the influence of shift of data on the classification performance as well as simple ways to mitigate these problems are also investigated. The results presented herein are also expected to be applicable in other areas of, e.g., spectroscopy, chomatography, mass spectrometry, etc., where large data

sets may slow down a "detection/warning" event or reducing the performance of a classification algorithm.

Briefly, LIBS is an atomic emission spectroscopy technique where the sample is converted to a plasma by a high energy laser pulse and optical emission spectroscopy is then performed as the plasma cools down, allowing to measure characteristic emission from the constituents of the sample. For more in depth theory, see e.g., Ref.[20] and references therein. In our set-up, an echelle type spectrograph was used to acquire the optical emission spectra, which resulted in that quite a lot of data (1 Mpixel camera in our case) needs to be read and transferred to a computer. By software conversion of the echellogram into an intensity vs. wavelength 1D spectrum, the $\sim 10^6$ data points are reduced to 4.8×10^4 ($\sim$5%). In addition, the use of full resolution images limits the overall sampling frequency in this particular set-up.

In the current work, LIBS spectra were acquired from pressed pellets of the different biomaterials, to increase the reproduceability of the single shot spectra and thus having better control of the input data used in the classification models. In a real life bioaerosol monitoring system, however, there will be a need to analyze single μm-sized particles[21] and using that data which in general will be of worse quality, compared to spectra from bulk material, in terms of, e.g., signal-to-noise, and a strong variation of relative line intensities is also expected[22].

Multivariate data analysis methods, such as Principal Component Analysis (PCA)[23,24] and Partial Least Squares (PLS) regression[25], were applied on the obtained spectral data sets, before and after various manipulation and variable reduction steps, and the results (mostly in terms of classification performance), are discussed and related to the data optimization.

Another common problem in, e.g., spectroscopy is a possible drift/shift of the captured data and such shifts usually rapidly degrades the performance of classification models. Here, we investigate the possibility to define windows (regions of interest), where important parameters are expected, and using the peak values from each window for building classification models. The measures of classification performance are compared relative to each other, i.e., no specific value is set as a threshold for any given model to be acceptable, as this will depend on the given application and the acceptable rates of errors (e.g., false alarms).

Finally, the present study only addresses ten substances, even though some very similar, and will of course be a much more favorable task compared to most real life scenarios where the samples of interest are likely to occur within a harmless background of complex composition.

II. EXPERIMENTAL

A. LIBS set-up

A pulsed (duration ∼5 ns and energy of 40 mJ) Nd:YAG laser (λ = 1064 nm) was focused to a spot diameter of about 300 μm onto pellets of different biomaterials (Dry samples were compressed into $d = 12.7$ mm pellets using a compacting pressure of ∼0.2 GPa. No binder was used.). The ablation was performed in standard ambient conditions and the laser induced plasma plume was imaged onto a $d = 550$ μm optical fiber connected to an echelle spectrograph equipped with an intensified CCD detector with a pixel resolution of 1024×1024 13.5 μm square sized pixels. Spectral data was analyzed either as echellograms (the intensity distribution directly from the CCD chip, see Figure 1) or from extracted 1D (intensity vs. wavelength) spectra. The 1D spectra were extracted, after initial spectral calibration by an HgAr lamp, from the echellograms by a transfer function which was performed by the control software (Andor Solis). Here, a wavelength window of ∼200-900 nm with a spectral resolution $\lambda/\Delta\lambda > 4000$ was obtained, and presented as ∼24000 data pairs. Converting an echellogram to 1D spectra may unfortunately introduce unwanted features, e.g., due to "cross-talk" between the diffraction orders. In addition, a stable temperature of the hardware, relative to calibration temperature, is of great importance as a change in only a few degrees will shift the position of the echellogram on the CCD chip because of temperature dependent optical properties of certain components. (Different algorithms are implemented in the software to reduce these errors.)

Compressed pellets, consisting of selected simulants for biological warfare agents and possible interferents (pollen) were used simply to improve the spectral reproducibility and quality, compared to, e.g., powder samples or single bioaerosol particles. To minimize the effects of drift in temperature and other experimental conditions, the whole data set was obtained during one trial. This data set contained observations from ten samples which each provided the data set with >100 observations, for a total of 1055 observations. To include an internal variation within a sample, the ∼100 observations were obtained from two different locations on the sample. The samples, which for convenience were numbered 1 through 10, were: three strains of *Bacillus atrophaeus* (BG), four strains of *Bacillus thuringiensis* (BT), *Betula pendula* (birch) pollen, *Pinus sylvestris* (pine) pollen, and Ovalbumin, see also Table I. The BG samples are essentially the same microorganism that have been cultivated and prepared in different ways. This is also the case for the BT samples.

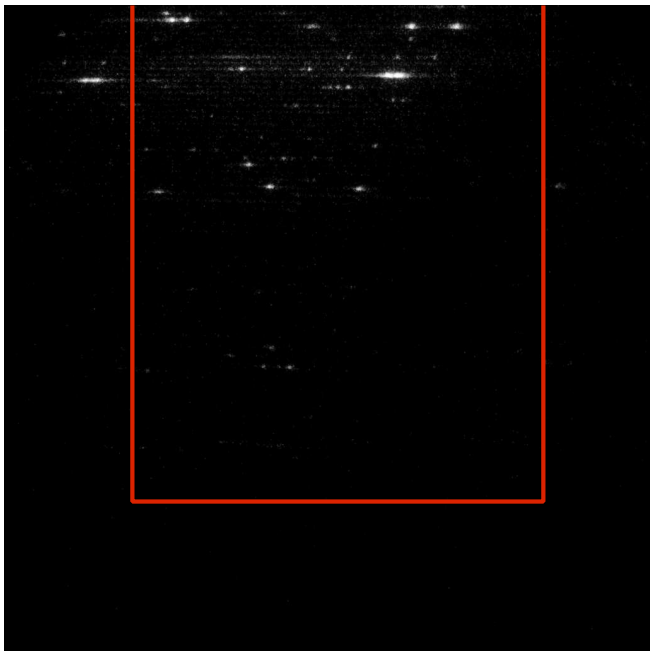

FIG. 1: Typical CCD image captured during LIBS on one of the bio-samples (BG New). The red lines show the approximate area of the image where the 1D spectra are extracted and where the images, in some analysis, were cropped.

TABLE I: Short description of the agents used in the analysis and their corresponding sample numbers used in plots.

Sample #	Name	Description
1	BG Nowoz	*Bacillus atrophaeus* dry spore preparation from Nowozymes
2	BG New	*Bacillus atrophaeus* dry spore preparation, newer batch
3	BG Old	*Bacillus atrophaeus* dry spore preparation, older batch
4	BT Turex	*Bacillus thuringiensis* dry spore preparation ”Turex WP” (commercial insecticide product)
5	BT Biobit	*Bacillus thuringiensis* dry spore preparation ”Biobit” (commercial insecticide product)
6	BT Bovall	*Bacillus thuringiensis* dry spore preparation by FOI
7	BT Dipel	*Bacillus thuringiensis* dry spore preparation ”Dipel” (commercial insecticide product)
8	Betula	*Betula Pendula* (birch) pollen (possible interferent)
9	Ovalbumin	Ovalbumin protein grade V, Sigma Aldrich (simulant for toxin)
10	Sylvestris	*Pinus Sylvestris* (pine) pollen (possible interferent)

B. Multivariate data analysis

From spectral measurements one may acquire a large amount of data and in a lot of applications the number of variables greatly outnumbers the observations, making multivariate data analysis a well suited method. One suitable method that can provide information if there are similarities between spectra is Principal Component Analysis (PCA). In addition, Partial Least Squares Discriminant Analysis (PLS-DA) was used to evaluate how different manipulation of 1D spectra and 2D echellograms affect the predictability compared to full resolution data.

1. Principal Component Analysis (PCA)

PCA is a projection method that transforms a data set onto a new variable space. The objective of the PCA algorithm is to find a linear transformation for the data to achieve a diagonal covariance (or correlation) matrix using eigenvalue decomposition. The components (eigenvectors) in the new variable space maximizes the variance in the original data set with as few variables as possible. These new variables are called principal components (PCs) and are constructed in such a way that the first component is put in the direction with the largest variance in the original data set, the second PC is put in the direction with the second largest amount of variance, under the constraint that it has to be orthogonal to the first PC, and so on. Mathematically, PCA can be written as

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{1}$$

where $\mathbf{X}$ is the original (measured) data matrix, $\mathbf{T}$ the scores and $\mathbf{P}$ the loadings. The scores matrix consists of the vectors which are referred to as the principal components and are the representation of the original data in principal component space. The loadings matrix contain the weights, or coefficients, which transform the data into principal component space under the restrictions mentioned above. The $\mathbf{E}$ matrix consists of error terms[26–28]. Typically, the resulting PCA is illustrated as a ”score plot” where data with similar properties will form a point swarm while other similar data will form other, separate point swarms. When using PCA to build models for prediction, it is common to include as many PCs that account for 80 or 90% of the variance in the original data set. (The cumulative variance is straightforward to calculate as the PCs are orthogonal and hence uncorrelated.) Any chosen level of explained variance will thus work as a threshold, where one assumes that the included PCs contain useful information from the signal and

the other (higher) PCs are assumed to contain mostly noise and should therefore be omitted. The number of components chosen are important (not only in PCA) in order to avoid a model which is over- or under-fitted, i.e., missing relevant information from too few components or introducing irrelevant information (noise) from too many components. In this study, PCA was only used for visualization purposes as it is a fast and simple approach to get information about how different data treatments affect the data set. For classification modeling purposes, there are more powerful methods, e.g., partial least squares regression to analyze and assess the data.

2. Partial Least Squares regression (PLS)

PLS is similar to PCA since it is built by a new set of components, or in this case referred to as factors, and it works especially well when the data has a lot of variables (i.e., when the number of variables is much larger than the number of observations). What makes PLS differ from PCA is that it is a supervised method, that is, there exist a response $\mathbf{Y}$ to the data matrix $\mathbf{X}$ where the general model is constructed as

$$\mathbf{Y} = \beta\mathbf{X} \tag{2}$$

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathbf{T}} + \mathbf{E} \tag{3}$$

$$\mathbf{Y} = \mathbf{U}\mathbf{Q}^{\mathbf{T}} + \mathbf{F} \tag{4}$$

where $\mathbf{T}$ and $\mathbf{U}$ are the scores and $\mathbf{P}$ and $\mathbf{Q}$ are the loadings of $\mathbf{X}$ and $\mathbf{Y}$ respectively. The matrices $\mathbf{E}$ and $\mathbf{F}$ are error terms. The objective of PLS is to find the matrix β to be able to predict new observations in $\mathbf{X}$. The PLS algorithm[26–28] uses the decompositions of $\mathbf{X}$ and $\mathbf{Y}$ to maximize the covariance between the score matrices $\mathbf{T}$ and $\mathbf{U}$. In this study, a special case of PLS was used, Partial Least Squares Discriminant Analysis (PLS-DA), which uses a binary response matrix $\mathbf{Y}$.

In a supervised model, one also has the benefit of getting a quality measure. In PLS the Q^2 value is a measure of how well a model can predict new data. This value is obtained from cross validation, i.e., a method where you create one calibration set and one test set from your data. The calibration set contains all but one observation which instead is put in the test set. The model is then built on the calibration set and tested by predicting the single observation from the test set. The procedure is repeated until all observations have been put in the test set once. The Q^2 value is defined as

$$Q^2 = 1 - \frac{\sum_i (y_i - \hat{y}_{(i)i})^2}{\sum_i (y_i - \bar{y})^2} \quad (5)$$

where the nominator is called PRESS (prediction residual sum of squares) where $\hat{y}_{(i)i}$ is the predicted value of y_i when that observation was excluded from the calibration set. The Q^2 value is also useful for minimizing the risk of over- or under fitting, e.g., by successively increase the number of factors and stop when no significant increase in predictability is observed[29]. In this study, a rule of significance was chosen in such a way that a factor is deemed significant if the increase in Q^2 from the previous factor is greater than one percentage point, i.e., $Q^2_{n+1} > Q^2_n + 0.01$. In the models, a Q^2 value was obtained for each of the ten samples and the total Q^2 value of a model was calculated as a mean of these individual predictabilities. Finally, to create a model and verify its predictability the commonly used splitting strategy of putting 2/3 of the observations in a training set was used. This fraction was found to be close to an optimal size to train models when a large data set (observations >100) is used[30]. In this investigation, the observations to be used in training and test set were randomly selected once and then used in all validations, in order to make better comparisons between models. From the validation, measures such as true/false positive/negative rates (and confusion matrices) could also be obtained and compared. No predefined values are given as thresholds for an acceptable predictor, instead the discussion will focus on their relative behavior.

C. Data processing and reduction of variables

As mentioned earlier, to obtain 1D spectra in our echelle spectrograph, the whole CCD image has to be read out at full resolution, after which the data is transferred to a computer where a software performs the extraction of the 1D spectrum. In this process, from a practical point of view, there would be little gain in overall sampling rate for our current experimental set-up to reduce the number of variables fed into a prediction model (except some improvement in computational time) if using these extracted 1D spectra for building prediction models. However, some studies were still performed on the extracted 1D spectra, such as reducing the amount of variables and normalization influence on predictability via PLS-DA. Here, the spectral data were normalized to unity. Instead of reducing the variables fed into the models via multivariate selection techniques[31], the simple assumption that useful information is found where peaks are observed in the spectra was used, i.e., the significant variables were reduced by an successively increased threshold using

the median of the standard deviation. Here, data were incrementally reduced by only including variables with standard deviation $s \geq 2^N \tilde{s}$, where $\tilde{s}$ is the median standard deviation and $N = 0, 1, 2, \ldots, 8$. In addition, reduction of variables was also performed by only include the maximum value inside 78 and 16 wavelength windows ($\Delta\lambda = \pm 0.26$ nm) centered around the strongest peaks found from normalized and averaged spectra.

As the step from image (or echellogram) to a 1D intensity vs. wavelength spectrum seems unnecessary for our purposes (especially as errors may be introduced), we also used PLS-DA models directly on the CCD data. Here, there are possibilities to significantly reduce data read-out and transfer times by binning the pixels and/or cropping the image. (Again, to further reduce the number of variables, pixels were chosen statistically as described above.) As the diffraction orders in the echellogram used to extract spectra are found in an area covering approximately half of the CCD area, it was also investigated if only the variables from that part were important for a model. Creating models using PLS-DA on different binned pixels, cropped images, variable reduction and comparing the Q^2 values results in a relative measure of how much is lost (if any) in predictability compared to what could be gained in data reduction. Similar as for the 1D spectra, using pixel windows and its largest pixel value was also studied on the images. This method was then applied to shifted images (which would simulate, e.g., thermal drift of the echellogram relative its calibration position). All image manipulation, such as cropping, binning and shifts, were performed by the Image Reduction and Analysis Facility (IRAF) software[32].

III. RESULTS AND DISCUSSION

Both the extracted 1D spectra and the echellograms (2D images) were analyzed via multivariate data methods. The main purpose was to see how different types of data reduction would affect prediction performance, as it may open up possibilities to reduce data read out and transfer times. Of course, in our particular set-up, a reduction of the number of variables in the 1D spectra will not reduce the acquisition time. However, analysis was performed on that type of data as it might be useful for other set-ups and applications. A shift in the data relative "calibrated" data was also introduced and its implication on classification performance was also studied.

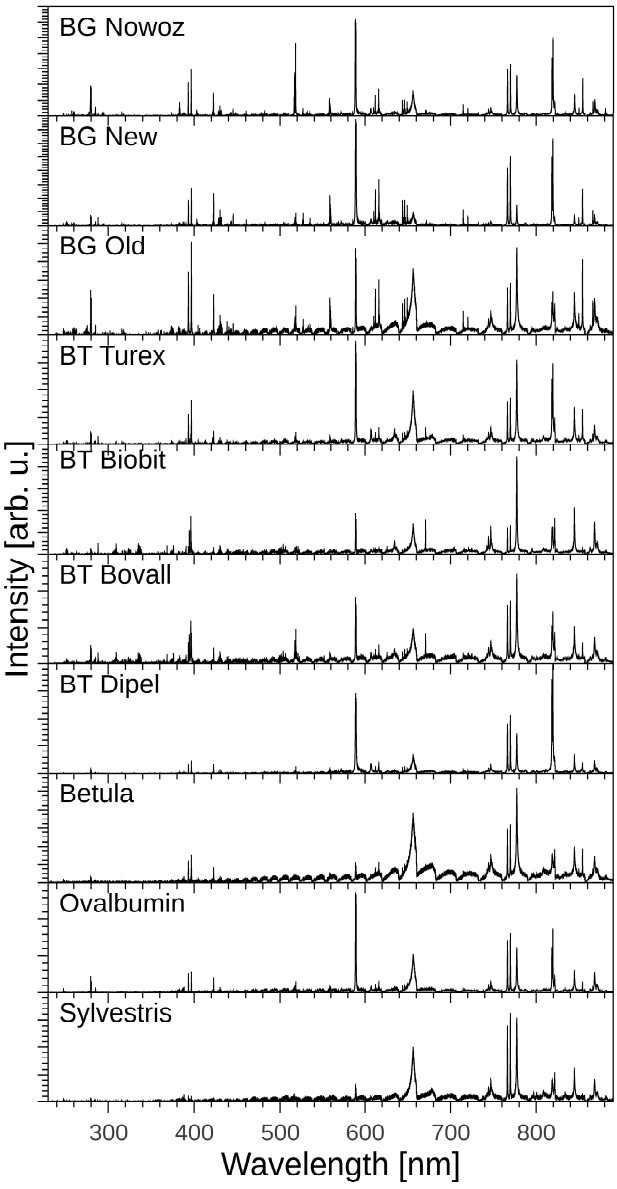

FIG. 2: Averaged LIBS spectra from the different samples, see Table I.

A. 1D intensity vs. wavelength spectra

Average spectra from the different bio-samples are shown in Figure 2. Even though the close similarities of some of the samples, see Table I, differences can be spotted by the naked eye which suggests that a prediction model could perform well (especially considering that, e.g, a PLS-DA model only has the ten different samples to consider). However, in a prediction model, single shot spectra will be used, where the signal-to-noise ratio is significantly reduced compared to the average from $\sim$100 spectra depicted in Figure 2.

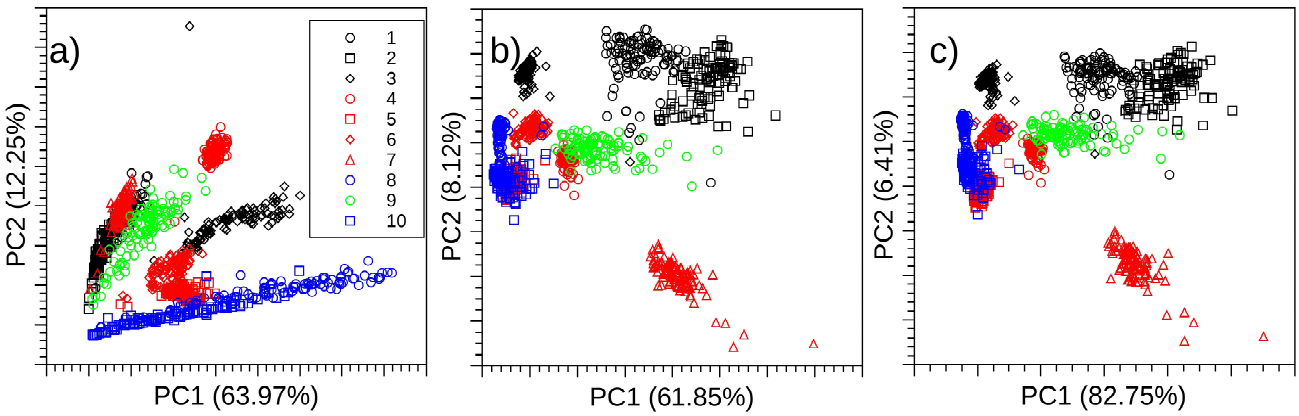

FIG. 3: Score plots of the first two principal components obtained from a) raw 1D spectra (∼24000 data points), b) 1D spectra normalized to unit area, and c) reduced number of variables to 78 (obtained from windows where the largest peaks were found). The percentage values on the axes are the explained variance.

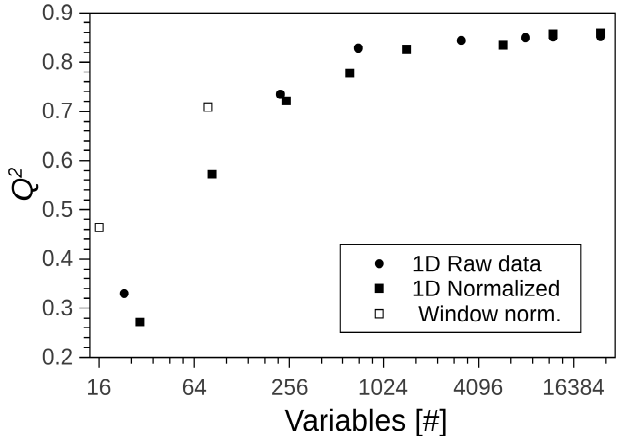

FIG. 4: Plot of Q^2 values obtained by PLS-DA of 1D spectra (raw and normalized) while reducing the number of variables (by changing the threshold, see subsection II C) and highest intensity value from windows (±0.26 nm) containing the strongest 78 and 16 peaks. In accordance to the rule of significance, 12 or 13 factors were used to create PLS-DA models on samples with more than ∼1000 variables and 4-13 factors when the number of variables < 1000. (Note the logarithmic scale on the x-axis.)

As a first step, the separability is visualized by PCA score-plots, see Figure 3. Here, as examples, score-plots from the untreated (raw single shot spectra), normalized spectra (to reduce the impact of intensity variations), and spectra where the number of variables have been reduced to only 78 by selecting the highest intensity peaks (after normalization) and defining spectral windows ($\Delta\lambda = \pm 0.26$ nm) from which the highest intensity value was taken. Each sample is labeled

with its unique color and marker combination, however, similar types of simulants have the same color in the plot. As can be seen, each sample creates a cluster (already when considering only the first two components) which shows that there are differences between samples but similarities in spectra within the same sample. When normalizing each spectrum to unity, the groups form smaller clusters, see Figure 3b. That is, the intensity variation between (and within) each sample group contributes to the variance in the first principal components and may thus affect the classification model. In Figure 3c, the number of input variables were dramatically reduced by selecting the largest value from normalized spectra within 78 wavelength windows where the strongest peaks over an arbitrary threshold was found when analyzing all spectra. Even this heavy reduction in input variables (from $\sim$24000 to 78) still show separate clusters. (As expected, this result confirms that the peaks in the LIBS spectra contain the important information.) Finally, the explained variance decreased after reducing the intensity variation (by normalization) and increased as the number of variables decreased (and only contained strongest peak data). Considering the larger sample subgroups (three BG samples, four BT, two pollen, and Ovalbumin) one might observe some grouping, even if only the first two components are visualized here, suggesting spectral similarities, however, no analysis in that direction will be presented in this work.

Results from PLS-DA prediction models of the 1D spectra are shown in Figure 4 where Q^2 values are plotted as function of number of variables. Only a small decrease in Q^2 is observed while reducing the number of input variables to $\sim$500. Below that value, however, a steep decrease can be seen. The number of factors used in the calibration models and cross validation to obtain the plotted Q^2 values were chosen in analogy with the pre-defined rule of significance (see subsection II B 2). Normalizing the 1D spectra does not influence the predictability much even though a quite large difference was observed in the PCA score plots, see Figure 3. Interestingly, a large increase in Q^2 was observed if the highest value inside predefined wavelength windows were used to create the model. For example, if only the 78 most intense peaks, within the predefined windows, were used in the PLS-DA, Q^2 increased from 0.57 to about 0.71 compared to selecting input variables by their variance as explained in subsection II C.

Another way to visualize the performance of the model is to make boxplots of values obtained from test and training sets using PLS models. Here, the model is built on the training set, randomly selected from 2/3 of the observations, and then tested with the remaining third of the observations and examples of the prediction can be seen in Figure 5. The red line inside the box corresponds to the median value of the predictions and the boundaries of the box are the 25th and 75th percentiles

TABLE II: Statistical data for the models depicted in Figure 5 containing mean, standard deviation (std), number of observations (N) and where mean±conf. creates a confidence interval for the observed values. The confidence interval is based on a student's t-distribution with a confidence level of 0.05 and N-1 degrees of freedom. Included are also the false negative rates (fnr), false positive rates (fpr), true positive rates (tpr), and true negative rates (tnr) of the two prediction models.

Positive predictions (Normalized 1D spectra, ∼24000 variables)											Positive predictions (Normalized 1D spectra, 78 variables from windows)									
Sample	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
mean	0.8981	0.9195	0.8907	0.8730	0.9401	0.8589	0.9495	0.7153	0.9033	0.7610	0.8746	0.8865	0.6136	0.6612	0.8851	0.6771	0.8841	0.4930	0.8098	0.7106
std	0.1721	0.1633	0.0724	0.0861	0.0865	0.0862	0.1798	0.2218	0.0784	0.1806	0.1778	0.1804	0.1700	0.1040	0.1086	0.0926	0.1650	0.1190	0.1745	0.2141
N	33	33	33	33	33	33	33	33	33	33	33	33	33	33	33	33	33	33	33	33
conf.	0.0610	0.0579	0.0257	0.0305	0.0307	0.0306	0.0638	0.0787	0.0278	0.0641	0.0631	0.0640	0.0603	0.0369	0.0385	0.0328	0.0585	0.0422	0.0619	0.0759
Negative predictions (Normalized 1D spectra, ∼24000 variables)											Negative predictions (Normalized 1D spectra, 78 variables from windows)									
Sample	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
mean	0.0044	0.0116	0.0114	0.0146	0.0003	0.0260	0.0069	0.0250	0.0126	0.0305	0.0064	0.0123	0.0360	0.0546	0.0033	0.0485	0.0126	0.0425	0.0234	0.03873
std	0.0331	0.0585	0.1048	0.1146	0.0887	0.1210	0.0629	0.1257	0.1125	0.1371	0.0476	0.0721	0.1734	0.1479	0.1156	0.1756	0.0823	0.1707	0.1335	0.1452
N	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297	297
conf.	0.0038	0.0067	0.0120	0.0131	0.0101	0.0138	0.0072	0.0143	0.0128	0.0157	0.0054	0.0082	0.0198	0.0169	0.0132	0.0200	0.0094	0.0195	0.0152	0.0166
Confusion matrix data (Normalized 1D spectra, ∼24000 variables)											Confusion matrix data (Normalized 1D spectra, 78 variables from windows)									
Sample	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10
fnr	0.0034	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0199	0.0000	0.0034	0.0067	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0199	0.0000	0.0034
fpr	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0357	0.0294	0.1579	0.0000	0.0000	0.0294	0.0000	0.0000	0.0000	0.0000	0.0357	0.0294	0.1579
tpr	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9643	0.9706	0.8421	1.0000	1.0000	0.9706	1.0000	1.0000	1.0000	1.0000	0.9643	0.9706	0.8421
tnr	0.9966	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9801	1.0000	0.9966	0.9933	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9801	1.0000	0.9966

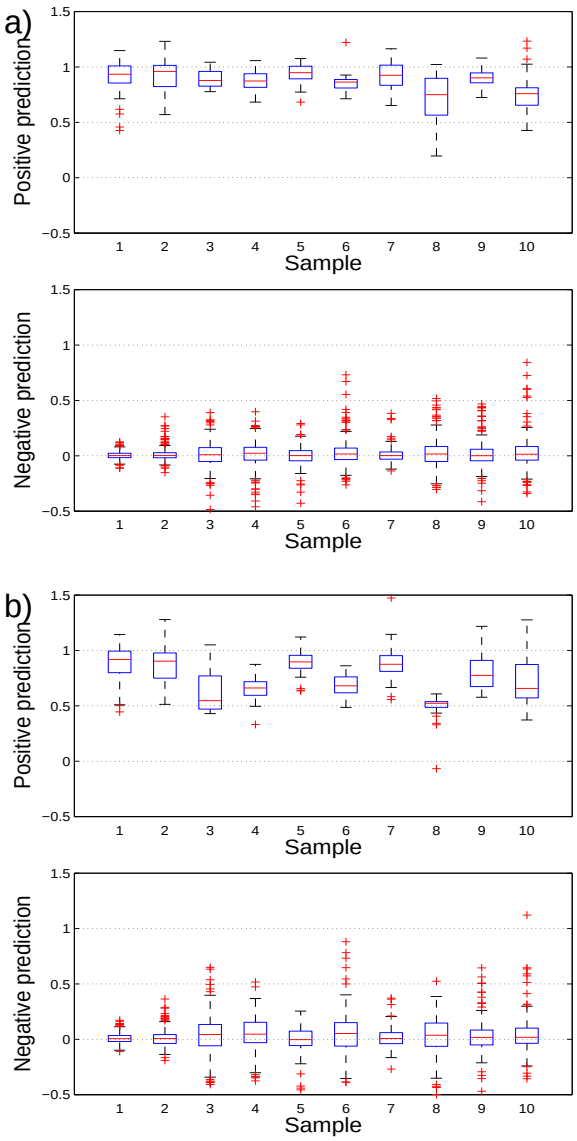

FIG. 5: Predictions from models using a) normalized full 1D spectra (∼24000 variables) with 13 factors and b) 78 variables with 11 factors. Ideal value for positive predictions is 1 and for negative 0. Red values are outliers (considered too extreme to come from the groups distribution). Statistical data from the predictions are presented in Table II

.

and will thus include 50% of the observations. The whiskers extend to the most extreme values considered not to be outliers by the algorithm and the outliers are plotted separately (in this case an outlier was defined as an observation lying outside $\pm 2.7\sigma$, where σ is the standard deviation of the distribution). Since PLS-DA is used, the ideal value of a true positive prediction is one and the

ideal value of a true negative prediction is zero, so preferably, the boxes in the different boxplots should be centered on these two values as tight as possible. As seen in Figure 5 (lower panels), the negative predictions are in general rather well centered around zero in every sample. The positive predictions, however, have a bit larger boxes and extended whiskers and for the model built on 78 variables and 11 factors according to the rule of significance (Figure 5b), the boxes have also moved down to lower values, compared to the model built on normalized full spectra, for most samples. However, of importance is that the boxes of the positive and negative predictions do not overlap, which they do not except maybe for sample 8. As mentioned earlier, Figure 5 is a visual representation of the performance of the model. To check statistically if the two prediction groups are separate and independent a two sample t-test was used, where the tested null hypothesis was that the positive and negative predictions come from distributions with equal mean. The test confirmed that the null hypothesis could be rejected and that positive and negative predictions are independent distributions at a 5% significance level, see also Table II.

The predicted results of the two models depicted in Figure 5 are shown in Table II. While both models are statistically separated on a 0.05 confidence level, it is also of interest to compare the different true/false positive/negative rates of the two models. Here, eight misclassifications (seven of them between the two pollen) were obtained using full spectra and only a slight increase could be seen using the 78 input variable model with nine misclassifications (seven between the two pollen).

In practice, another problem commonly occurs. When performing calibration of your system, there is often a change in the x-axis. One way of compensating for these changes is to introduce an interpolation step for all spectra ('database' and new measurements) to make sure that a common x-scale is used. Additional smoothing of the data by, e.g., a Savitzky-Golay filter may also be performed. (The interpolation step (and filter) commonly also reduce the number of variables.) However, in our case, introducing an interpolation only marginally reduced the Q^2 values and applying a smoothing filter on the data only decreases the Q^2 values for heavy variable reduction (if only variables with large variance are included), i.e., reducing the number of variables to < 500, results not shown.

To explore the effect of spectral shift on the prediction models, spectra were shifted in x (wavelength scale) by 1, 2, 4 and 8 data point steps up and down while testing them in a model constructed from non shifted data. The true negative and positive rates for the shifted spectra can be seen in Figure 6. For the full resolution spectra (black symbols), a shift in maximum two steps only

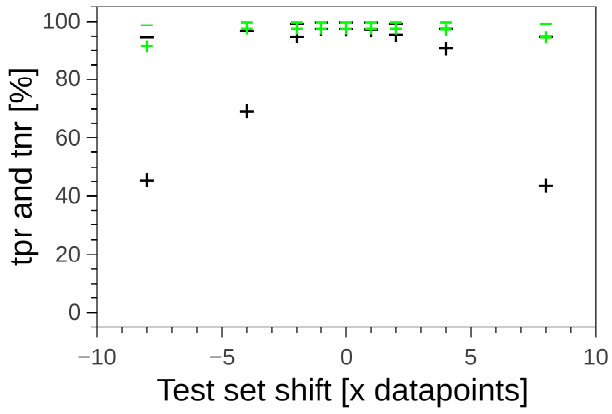

FIG. 6: True positive rate (tpr, +) and true negative rate (tpr, -) for the PLS-DA models on 1D spectra as the test set is shifted relative the calibration set. The black symbols show the results from the full (∼24000 variables) spectra and green symbols the results from model created by the 78 wavelength windows.

affects the model slightly, while a larger decrease was found for four steps and above. However, if using the highest value found in the predefined 78 wavelength windows (where only a minor decrease in misclassifications were found compared to the full 1D spectra, see Table II), a larger shift can be tolerated without losing much in predictability (green symbols).

B. Images

For the images, a similar approach as for the 1D spectra was used. That is, the full resolution images were used to build a prediction model and compared to models built on statistically reduced number of variables. Images were also binned in different ways to further reduce the image data size. The effect of cropping the images to only include the region from which the 1D spectra were extracted (see Figure 1) was also studied. In addition, the impact of a shift in x or y of the echellogram (where the shift in y simulates a hardware temperature drift in our case) was also studied.

Q^2 values as function of number of variables for the different binned full and cropped images are shown in Figure 7. No large changes in Q^2 was observed when using the cropped (approximately half-size) images when comparing to the full size images and similar number of variables. As the peaks/spots outside the main echellogram area (from where the 1D spectra are extracted) originates from additional diffraction orders and will thus be of redundant character, the similar Q^2 values should then be expected for the cropped images. For both the image sizes, the Q^2 value

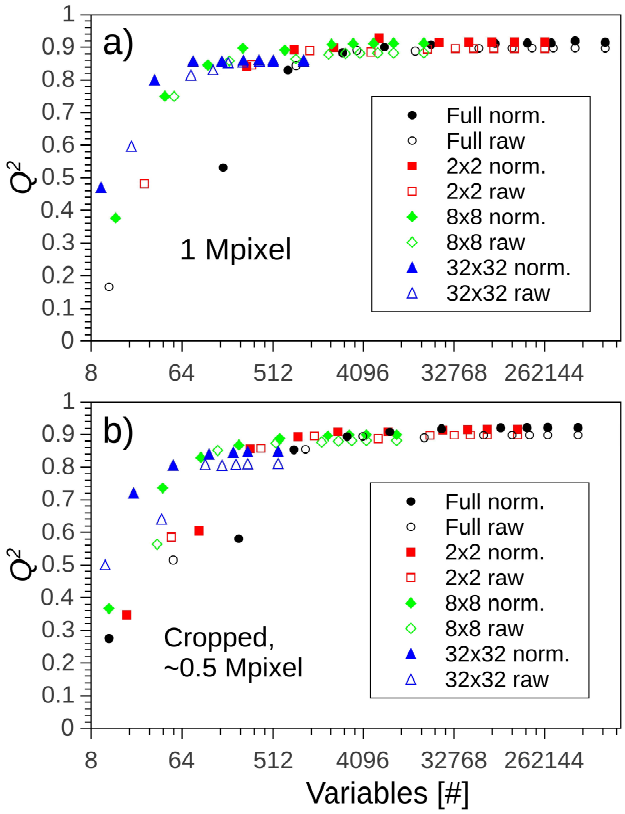

FIG. 7: Q^2 values obtained from PLS-DA on the CCD images (raw and normalized data). a) depicts the results using the full 1 Mpixel image and b) when only using approximately half the CCD data, i.e., the part used for 1D spectra extraction (see Figure 1). Note the log scale on the x-axis.

only decreases slightly when binning the pixels and reducing the number of variables to ∼100. Below ∼100 variables, however, a steep decrease occurs. Compared to the 1D spectra, however, see Figure 4, one can reduce the amount of variables more while still obtaining a relatively large Q^2 and there are also higher Q^2 values for the image data at the same amount of variables. As the absolute values of Q^2 decreases (at similar amounts of variables) and that the steep decrease is observed at a larger number of variables for the 1D spectral data (Figure 4), it clearly indicates that there is information lost when converting the echellograms into intensity vs. wavelength spectra.

In general, binning the images reduces the maximum Q^2 value (at similar number of variables), however, when reducing the variables using the thresholds in variance, larger Q^2 values are found for the binned images at less variables. Normalizing the images slightly improves the Q^2 value, especially for heavy binning (e.g., 32x32).

Boxplots of two prediction models are depicted in Figure 8 and can be compared with the

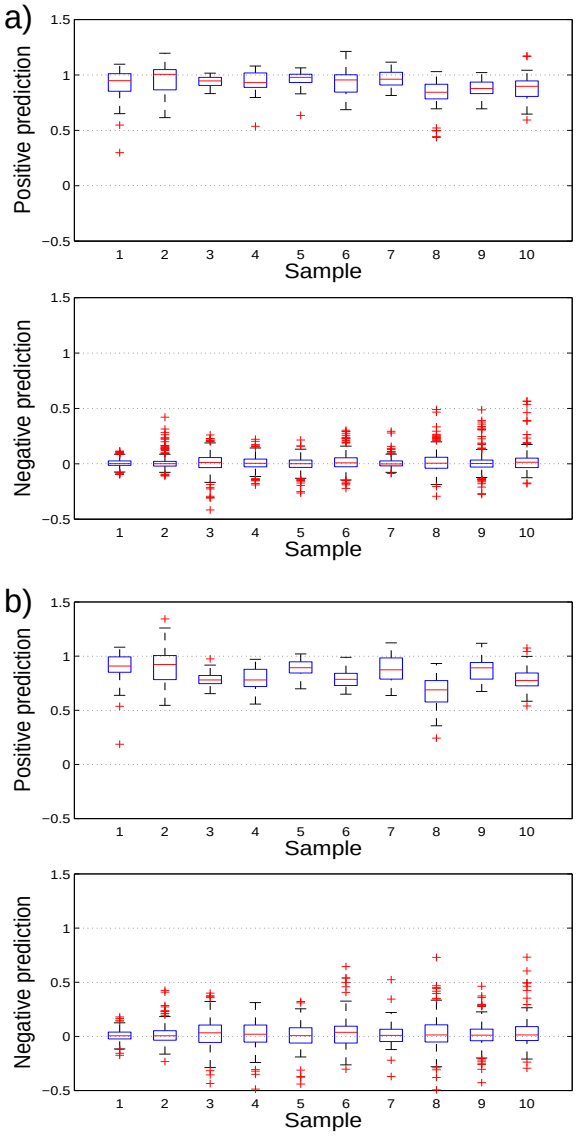

FIG. 8: Predictions from models using normalized a) cropped full resolution images (echellograms), ∼560000 variables and 12 factors, and b) the 51 variables with largest variance from cropped images binned 32x32 and 13 factors. Ideal values for positive predictions is 1 and for negative 0. Red values are outliers (considered too extreme to come from the groups distribution).

results presented in Figure 5. Here, the result for a prediction model based on the full resolution, cropped images are shown in Figure 8a, where tighter boxes closer to the ideal values can be seen compared to the full resolution 1D spectra (see Figure 4a), again illustrating the advantage of using images instead of the extracted spectra. As an example of binned and reduced variables, the result from a model created from cropped images binned 32x32 and only using the 51 variables

with largest variance is shown in Figure 8b. Such heavy reduction of the image data still results in separate distributions of the classifications. For the full resolution, but cropped, echellograms ($Q^2 = 0.92$ using 12 factors), a total of four observations were misclassified (three between the pollen) and by using only the 51 largest variance variables from the 32x32 binned and cropped images ($Q^2 = 0.81$ using 13 factors) a total of six observations were misclassified (five between the pollen). Again, these results illustrate the better performance of a classification algorithm built on data from the echellograms as compared to extracting the more commonly used 1D spectra. For our purposes, i.e., to speed up readout and data transfer times without loss of predictability, these results are indeed very helpful. In this case, instead of having to read out the whole full resolution CCD image and perform a conversion/extraction of a 1D spectra it is in some cases enough (or even better) to only read out half the CCD area and also bin the pixels to some extent before building, or feeding data into, a prediction model. It should also be noted, however, that the prediction accuracy (in terms of, e.g., tnr and tpr) is not directly proportional to the Q^2 values. That is, the randomly chosen images to build and validate the model may be more favorable for a model corresponding to a lower Q^2. In addition, there might still be an influence of over- or underfitting of the models by choosing the number of factors by the rule of significance used here. For example, using the normalized cropped and 32x32 binned echellograms (572 variables and model based on 13 factors with $Q^2 = 0.85$) results in only two misclassifications and using normalized cropped images binned 1x8 (70296 variables and model based on 13 factors with $Q^2 = 0.90$) results in 100% correct classification for the randomly selected spectra used in this study.

Finally, the impact of shift in image input data was also investigated. The cropped images were shifted in x- and y-directions and a model made from the original (not shifted) images was used to predict these shifted images and Figure 9 shows the tpr and tnr obtained. As can be seen, the model is more sensitive to a shift in y-direction (which corresponds to the thermal drift direction), where already a one pixel shift significantly reduces the tpr for the full resolution images. Likely, this behavior is because of the asymmetric distribution of a spot in the echellogram, i.e., an elliptic intensity distribution can be seen (especially at shorter wavelengths where the pixel resolution in pixels/nm is higher) with the large axis in the x direction. One way to reduce the problem caused by image shift could be to bin the images before creating a model and test set. In Figure 9b it can be seen that the model becomes slightly more robust and almost similar values are obtained for one pixel shift in any direction (Note that in this case the number of variables also decreases to 572 compared to ∼520k for the full resolution cropped images). In addition, and similar to the

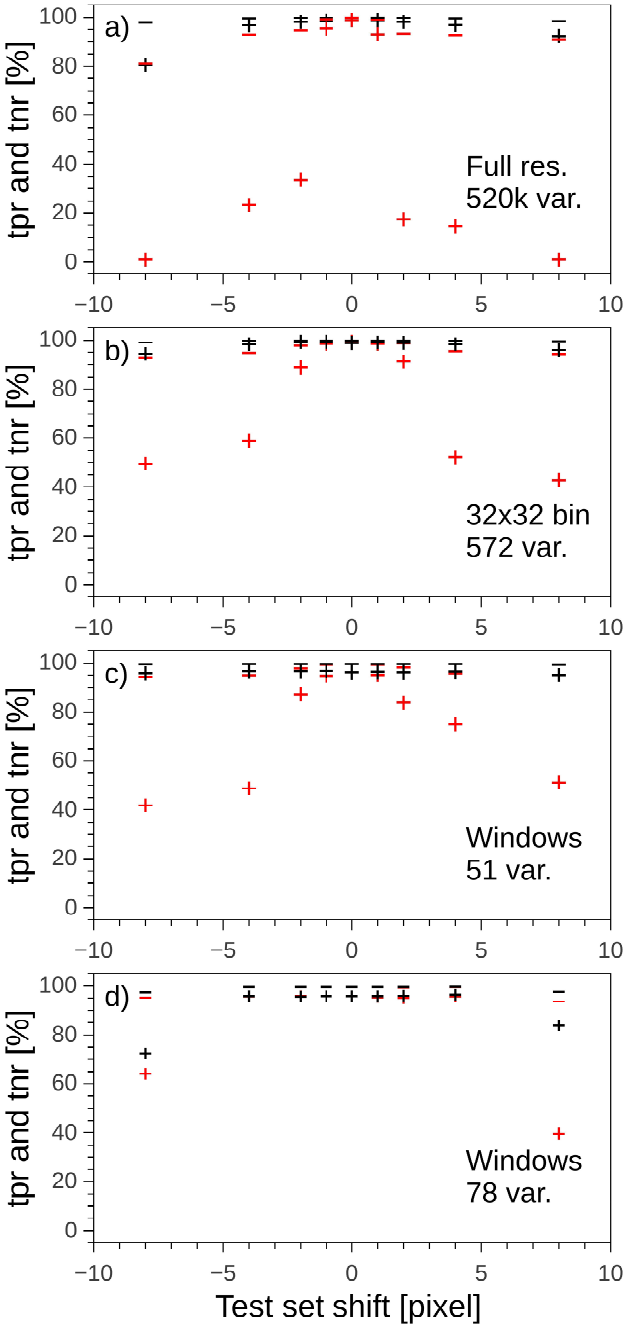

FIG. 9: True negative rate (tnr) and true positive rate (tpr) from shifted cropped images. a) cropped full resolution images (echellograms), ∼560000 variables and 12 factors, b) cropped echellograms binned 32x32 (572 variables and 13 factors), and c) windows. tpr are marked by (+) and tnr by (-) in the plot where black symbols corresponds to an x-shift and red a y-shift.

1D spectra, regions of interest were defined and the largest pixel value was then found in each window and used to build a prediction model. As an example, using 51 32x32 pixel windows (regions selected by largest variance in 32x32 binned images) and finding the largest value also seem to improve the robustness of the model, see Figure 9c. Here, however, only 51 variables were used resulting in slightly lower tpr and tnr rates as compared to the models built with more variables (Figure 9a and b).

IV. SUMMARY

The study of both the extracted 1D spectra and images showed that the LIBS data of the different biosamples can be quite well separated using multivariate data analysis methods. The PCA of the spectra and images showed that differences between samples were observable and that normalized data showed even more distinct separation of the samples while illustrating only the first two components. However, for 1D spectra the predictability of the models showed little or no improvement with normalized data whereas for the images a slight increase was observed (especially for the heavily binned images). From the analysis of single shot 1D LIBS spectra it was found that a threshold can be used to reduce the number of wavelengths used for classification to about 500 without losing a significant amount of predictability (here represented by the Q^2 value). Using wavelength windows centered around the strongest peaks in the spectra created more robust models if introducing a spectral shift in the test set.

Using PLS-DA on the images as input resulted in an improvement in predictability compared to the spectra. In addition, only the region containing the main echellogram is needed when constructing the models. The images initially have more variables than the spectra, however, the current work shows that it is possible to reduce the number of pixels to orders of magnitudes less than the number of data points in a 1D spectrum and still obtain a better predictability.

Reducing the data size by decreasing the resolution of the images, i.e., by binning the pixels, also proved to be possible with little change in, e.g., tpr and tpr rates. Even a quite heavy binning (e.g., 32x32) can be performed without significantly affecting the predictability (compared to using full resolution images).

The simulation of, e.g., thermal shifts of the image revealed the fragility of PLS-DA models and a higher sensitivity was observed for shifts in the y direction. To slightly improve the robustness, methods such as binning and/or introduce windows (regions of interest) from which the largest

value is taken could be used.

Acknowledgments

This work was funded by the Swedish Department of Defence, Project no. 440-A404114 and a Grant from the European Union Infrastructure Committee for the project "Two Stage Rapid Biological Surveillance and Alarm System for Airborne Threats (TWOBIAS)", Grant no. FP7-242297. The authors would also like to thank Dr. T. Tjärnhage for fruitful discussions.

---

* Electronic address: lars.landstrom@foi.se

[1] P. Baldi, S. Brunak, Y. Chauvin, C. A. F. Andersen, and F. Nielsen, Bioinformatics **16**, 412 (2000).

[2] Y. Saeys, I. Inza, and P. Larranaga, Bioinformatics **23**, 2507 (2007).

[3] P. J. Lisboa, Neural Networks **15**, 11 (2002).

[4] H. Lohninger and K. Varmuza, Anal. Chem. **59**, 236 (1987).

[5] P. Jonsson, A. I. Johansson, J. Gullberg, J. Trygg, J. A, B. Grung, S. Marklund, M. Sjöström, H. Antti, and T. Moritz, Anal. Chem. **77**, 5635 (2005).

[6] P. Jonsson, G. Olofsson, and T. Tjärnhage, eds., *Bioaerosol Detection Technologies* (Springer-Verlag, New York, 2014).

[7] R. DeFreez, Proc. SPIE **7484:74840H**, 1 (2009).

[8] V. Sivaprakasam, A. L. Huston, C. Scotto, and J. D. Eversole, Optics Express **12**, 4457 (2004).

[9] J. D. Hybl, S. M. Tysk, S. R. Berry, and M. P. Jordan, Appl. Opt. **45**, 8806 (2006).

[10] C. D. Clark, P. Campuzano-Jost, D. S. Covert, R. C. Richter, H. Maring, A. J. Hynes, and E. S. Saltzman, J. Aerosol Sci. **32**, 765 (2001).

[11] H. Felix-Rivera and S. Hernandez-Rivera, Sens. Imaging **13**, 1 (2012).

[12] A. Sengupta, N. Brar, and E. J. Davis, Journal of Colloid and Interface Scienc **309**, 36 (2007).

[13] D. W. Hahn and M. M. Lunden, Aerosol Sci. Technol. **33**, 30 (2000).

[14] J. D. Hybl, G. A. Lithgow, and S. G. Buckley, Appl. Spectrosc. **57**, 1207 (3002).

[15] S. J. Rehse, Q. I. Mohaidat, and S. Palchaudhuri, Appl. Opt. **49**, C27 (2010).

[16] F. C. DeLucia and J. L. Gottfried, Appl. Opt. **51**, B83 (2012).

[17] A. C. Samuels, F. C. DeLucia, K. L. McNesby, and A. W. Miziolek, Appl. Opt. **42**, 6205 (2003).

[18] C. A. Munson, F. C. DeLucia, T. Piehler, K. L. McNesby, and A. W. Miziolek, Spectrochim. Acta B **60**, 1217 (2005).

[19] N. L. Lanza, R. C. Wiens, S. M. Clegg, A. M. Ollila, S. D. Humphries, H. E. Newsom, and J. E. Barefield, Appl. Opt. **49**, C211 (2010).

[20] A. W. Miziolek, V. Palleschi, and I. Schechter, eds., *Laser Induced Breakdown Spectroscopy* (Cambridge University Press, Cambridge, 2006).

[21] T. Tjärnhage, P.-Å. Gradmark, A. Larsson, A. Mohammed, L. Landström, E. Sagerfors, P. Jonsson, F. Kullander, and M. Andersson, Opt. Commun. **296**, 106 (2013).

[22] G. A. Lithgow and S. G. Buckley, Appl. Phys. Lett. **87**, 011501 (2005).

[23] K. Pearson, Philosophical Magazine **2**, 559 (1901).

[24] H. Hotelling, Journal of Educational Psychology **24**, 417 (1933).

[25] S. Wold, M. Sjöström, and L. Eriksson, Chemometrics and Intelligent Laboratory System **58**, 109 (2001).

[26] P. Geladi and B. R. Kowalski, Analytica Chimica Acta **185**, 1 (1986).

[27] S. Wold, K. Esbensen, and P. Geladi, Chemometrics and Intelligent Laboratory Systems **2**, 37 (1987).

[28] T. Rajalahti and O. M. Kvalheim, International Journal of Pharmaceutics **417**, 280 (2011).

[29] P. Geladi, B. Sethson, J. Nystrom, T. Lillhonga, T. Lestander, and J. Burger, Spectrochim. Acta Part B **59**, 1347 (2004).

[30] K. Dobbin and R. Simon, BMC Medical Genomics **4**, 31 (2011).

[31] T. Mehmood, K. H. Liland, L. Snipen, and S. Sæbø, Chemometrics and Intelligent Laboratory Systems **118**, 62 (2012).

[32] `http://iraf.noao.edu/`, *Accessed August 2014*.