Uncertainty assessment in FT-IR spectroscopy based bacteria classification models

Ornella Preisner a,⁎, João A. Lopes b, José C. Menezes a

a Institute for Biotechnology and Bioengineering, Centre for Biological and Chemical Engineering, IST, Technical University of Lisbon, Av. Rovisco Pais, P–1049-001 Lisbon, Portugal

b REQUIMTE, Serviço de Química-Física, Faculdade de Farmácia, Universidade do Porto, Rua Aníbal Cunha, 164, 4059-030 Porto, Portugal

A R T I C L E   I N F O

Article history:
Received 26 March 2008
Received in revised form 14 June 2008
Accepted 18 June 2008
Available online 22 June 2008

Keywords:
FT-IR spectroscopy
Microorganisms’ classification
Uncertainty estimation
Jackknifing
Bootstrapping

A B S T R A C T

The number of Fourier Transform infrared spectroscopy (FT-IR) applications for discrimination between microorganisms at different levels has increased substantially over recent years. Appropriate spectra handling and processing algorithms is a requirement for successful method implementation. Chemometrics methods such as Principal Component Analysis (PCA), discriminant Partial Least-Squares Regression (PLS-DA), and Soft Independent Modelling of Class Analogy (SIMCA) have found particular relevance in this context. Despite its applicability, robustness of multivariate calibration is still a key issue for effective application of FT-IR spectroscopy for microorganisms classification. Resampling methods provide a very interesting alternative to estimate dispersion measurements in the context of PCA, PLS-DA and SIMCA modelling. This work focuses on the comparison of PCA, PLS-DA and SIMCA approaches for bacteria discrimination based on FT-IR spectra. Bias and uncertainty for all implementations in terms of bacteria classification prediction are assessed using two non-parametric resampling methods — jackknife and bootstrap. A total of 73 samples of Acinetobacter baumannii, Enterococcus faecium and Staphylococcus aureus/epidermidis were used for calibration, and 32 samples from an independent samples group were used for model testing. Resampling strategies were applied to each method to provide a dispersion measure of the classifications for the testing set. Employed bootstrapping and jackknifing methods demonstrated to be valid alternatives to estimate bias and variance for non-supervised and supervised microorganisms discrimination models.

© 2008 Elsevier B.V. All rights reserved.

1. Introduction

There is a continuing requirement for accurate, rapid, and reliable methods for the discrimination among a wide range of clinically relevant pathogens and microbial strains. In the last 10 years, Fourier transform infrared (FT-IR) spectroscopy has been used as a complementary technique for bacteria classification and identification due to its rapid and highly specific ‘fingerprinting’ capabilities [1,2]. FT-IR spectroscopy generates large amounts of data, hence requiring appropriate multivariate statistical methods depending on the specific objective. Despite the existence of many successful applications of bacteria characterization with FT-IR in the literature, reports of successful field implementation of FT-IR spectroscopy as a routine basis hardly exist [3,4]. This is mainly due to the lack of robustness of the developed models. Robust estimates of the accuracy and precision of regression models are crucial for any method to be adopted in routine [5]. There are several statistics for measuring the accuracy of a model, such as the root mean squared error (RMSE) which represents the expectation of the squared differences between true and predicted values [6]. The RMSE has been extensively used to estimate model bias in chemometrics models [6,7]. Cross-validation and external validation strategies are commonly adopted to estimate model bias [7]. Often, only bias is used to optimize and evaluate the performance of chemometrics models. Giving the importance of uncertainty estimation in multivariate data analysis, different uncertainty-based approaches in chemometrics have been proposed [8]. Uncertainty is defined as a parameter associated with the result of a measurement, which characterizes the dispersion of the values that could reasonably be attributed to the measurement [9]. Basically, there are two ways of estimating uncertainty. Error propagation is a very convenient strategy mostly employed in univariate calibration. Resampling methods, instead, have the attractive property of requiring very little modelling and data properties assumptions, still maintaining the advantages of conceptual simplicity and accuracy. Westerhuis et al. (2008) described the use of permutation testing and cross PLS-DA model validation to assess the validation of classification models [10]. Jackknifing (as proposed by Quenouille (1956)) and bootstraping (as proposed by Efron (1970)) are as well the two resampling techniques with greater utilization among the chemometrics community [11,12]. These techniques have been extensively described in the literature, namely in statistics journals [13]. The attractiveness of jackknifing and bootstrapping is that they provide a significant and previously unattainable type of information: estimates of dispersion for statistics of unknown or poorly known distribution [14]. The importance of dispersion measures is reflected by the fact that in the analytical...
literature, sample means should always be reported with accompanying measures of dispersion as stated in the ICH-Q2 guideline [15]. In practice this is seldom the case, namely in chemometrics applied works. Generally, the function of jackknifing and bootstrapping is to generate the empirical distribution for a given statistic that permits the calculation of central and dispersion estimates (e.g., mean and standard deviation) by resampling the original data set [16]. Central or dispersion measures can be derived from the populations and used to define for example confidence or prediction intervals [17]. The jackknife often works well, provided that the statistic under study does not change drastically upon small changes in the data [18,19].

Previous studies on discrimination, classification, and identification of microorganisms applied both unsupervised and supervised techniques for representation of information from hyper-spectral data, namely Principal Component Analysis (PCA), Self-Organising Maps (SOM), Linear Discriminant Analysis (LDA), Hierarchical Cluster Analysis (HCA), Partial Least Squares Regression (PLS) and Artificial Neural Networks (ANN) among others [20–22]. However, in order to develop a robust model, one should assess the long term accuracy and robustness of the developed methods as well as the models precision.

The main purpose of this study was to evaluate the implementation of different resampling methodologies to assess the uncertainty of chemometrics methods for FT-IR-based bacteria discrimination models. Relative merits of the jackknife and bootstrap approaches for bias and uncertainty estimation in both supervised and unsupervised techniques such as PCA, discriminant Partial Least Squares (PLSDA) and Soft Independent Modelling of Class Analogy (SIMCA) are reported for a problem of pathogenic bacteria discrimination based on FT-IR data.

2. Theory

2.1. Non-parametric techniques

Two non-parametric resampling approaches are used here to analyze the uncertainty in model estimation. Locational (e.g., mean) and dispersion (e.g., standard deviation) statistics are extracted from the parameters obtained using the jackknife and bootstrap procedures.

2.1.1. Jackknife

The jackknife is an unbiased estimation strategy often described in the cross-validation context as the leave-one-out cross-validation strategy [23]. One of its major applications is to optimize the number of latent variables in PCA or PLS. The jackknife applied to a data matrix with \( n \) samples and \( k \) variables, builds \( n \) different matrices each one leaving one sample out. The structure of the resampled matrices is represented below.

\[
X_i = \{x_{1i}, \ldots, x_{ki}, x_{i+1}, \ldots, x_n\}, \quad i = 1, \ldots, n
\]  

(1)

If jackknife is to be applied in regression models, then the structure should include two arrays (the independent \( X \) and the dependent \( Y \) matrices).

\[
(X_i, Y_i) = \{(x_{1i}, y_{1i}), \ldots, (x_{ki}, y_{ki}), (x_{i+1}, y_{i+1}), \ldots, (x_n, y_n)\}, \quad i = 1, \ldots, n
\]  

(2)

The jackknife method always generates the same number of matrices as the number of samples.

2.1.2. Bootstrapping

Bootstrapping methods can be used to estimate empirical distributions of statistical parameters of interest (such as model coefficients or sample predictions). The bootstrap exists in a non-parametric and a parametric version [24]. There exist a large number of different variants of bootstrap: the residuals bootstrap, the 0.632 bootstrap, the half bootstrap, the wild bootstrap, the smoothed bootstrap, the m-out-of-n-bootstrap, the iterated bootstrap, the balanced bootstrap among all [18]. In this work we used the naïve version of bootstrapping.

As described by Wehrens et al. (2000) [18], bootstrapping can even be used to assess precision of confidence or prediction intervals. For each bootstrap a total of \( n \) samples are sampled randomly with replacement from the original data set \( X \). Because the sampling is performed with replacement, it is normal to have repetitions of samples in the same bootstrapped data set.

\[
X_i = \{x_{1i}, \ldots, x_{ni}\}, \quad i = 1, \ldots, B
\]  

If bootstrapping is to be applied in regression models, then the structure should also include two arrays.

\[
(X_i, Y_i) = \{(x_{1i}, y_{1i}), \ldots, (x_{ni}, y_{ni})\}, \quad i = 1, \ldots, B
\]  

(4)

The total number of bootstraps (\( B \)) should be large enough to ensure precision of the statistical parameters that are to be estimated. It is common practice to use more than 1000 and less than 5000 bootstraps in most situations. However, this number might depend on the exact purpose of the method and of the type and size of data. The more convenient way to define the number of bootstraps is to evaluate the effect of \( B \) on the estimation of the parameters. This can be done by repeating several times the bootstrap procedure and then to evaluate the standard deviation of the estimated statistical parameter. This practice however may turn out to be impractical due to the time required if data sets are large. It should be noted that the bias and standard deviation obtained by jackknifing and bootstrapping are different as described in Wehrens et al. (2000). For jackknife to be compared with bootstrap, jackknife bias should be multiplied by \((n-1)\) and standard deviation by \((n-1)^{0.5}\), where \( n \) is the number of samples.

2.2. Models

In this work one unsupervised (PCA) and two supervised models (SIMCA and PLSDA) were compared.

2.2.1. Principal Component Analysis (PCA)

PCA is a commonly used method to explore multivariate spectroscopy data [6]. Similarity between objects is normally assessed by plotting the scores of relevant principal components in scatter plots. Considering that the quality of data is good enough, for very large data sets the number of model calibration samples ensures robustness, while this cannot be expected for relatively small data sets. In this case, PCA models can be very dependent on the samples selected for calibration. Often, the objective of building PCA models is the identification of groups of samples based on scores clustering. It is however very difficult to define what a cluster is and the identification based on the visualization can be highly misleading. There are several automatic methods to identify clusters but these are always based on some clustering algorithm and often give no consistent results [25].

The jackknife and bootstrap resampling approaches can be applied for scores uncertainty estimation in PCA models [25]. For PCA the simple data resampling is not enough because scores obtained from these models might not be directly comparable due to the rotational ambiguity of PCA. Therefore, before comparing scores obtained from different models, the latter must be rotated so that they can match. The proposed strategy is to rotate the loadings of each resampled model so that they match a reference loadings matrix. Scores obtained from these rotated models are then compared and statistics like the mean or standard deviation can be obtained for each sample. The algorithm for loadings rotation can be found elsewhere [26]. Based on a calibration set, bootstrapping was used to identify different PCA models. The testing data set was then projected onto each bootstrapped model yielding for each testing sample a scores matrix with dimensions \( B \times k \) where \( B \) is the number of bootstraps and \( k \) is the
number of principal components. For each testing sample the standard deviation of the scores was calculated. This parameter reflects the precision of the scores estimation for each sample.

The selection of the number of significant number of principal components in PCA models can be performed in many ways [7]. Bootstrapping was also proposed here as a means to estimate the optimal number of components. An external testing data set was projected onto each bootstrapped model yielding for each testing sample a scores matrix with dimensions $B \times k$ where $B$ is the number of bootstraps and $k$ is the number of principal components. For each testing sample the standard deviation of the scores was calculated. This parameter reflects the precision of the scores estimation for each sample.

2.2.2. Soft Independent Modelling of Class Analogy (SIMCA)

SIMCA is based on the identification of local models built from defined groups. It aims at predicting a class membership for each observation [7]. In SIMCA, a PCA is performed on the data corresponding to each class in the data set. A sufficient number of principal components are retained to account for most of the variation within each class. Hence, a principal component model is used to represent each class in the data set. New samples can be projected onto the developed models and the class membership is evaluated according to the distance to these models. A measure of that distance can be based on the Hotelling ($T^2$) and squared prediction error (SPE) statistics [7]. These statistics measure the distance of the projected sample within the model and the residuals. Since these statistics are relative, normalization is required. It can be achieved by dividing the statistics by their corresponding confidence limit at a certain probability level ($T_{df}^2$ and SPE$_{\alpha}$). For each new sample $i$ projected on the PCA models for classes 1 to $J$, a vector of distances to each model is computed ($D_i$).

\[
D_i = (D_{i1}, \ldots, D_{iJ})
\]

where $D_{ij}$ is the distance of the sample to the model for class $j$.

The measure of the distance corresponding to sample $i$ for model $j$ is calculated according to Eq. (6).

\[
D_{ij} = \sqrt{\frac{\chi^2_{df_j, 1} \times \text{SPE}_{ij}}{\chi^2_{df_j, 1} \times \text{SPE}_{\alpha_j}}}
\]

(6)

The class membership for sample $i$ is then depicted according to the $D_{ij}$ values. Sample $i$ is assigned to the class with the minimum distance $D_{ij}$. A probability of class membership can also be estimated based on these distances [6].

2.2.3. Discriminant PLS (PLSDA)

PLS has been thoroughly used for classification purposes as a supervised modelling type. In the special case of classification models the dependent variables are classes, a particular kind of coding is required. Partial least squares targets (or classes) were codified using a vector of zeros and ones with the dimension equal to the maximum number of classes as described by Martens and Næs [7]. Therefore, $n$ samples belonging to $m$ different classes yield a binary data matrix with dimensions $n \times m$. A PLS model used to regress an independent variable vector $X$ to $m$ dependent variables yields a $m$-vector of predictions for model $i$ (PLSDA) [27]. By projecting a new sample $i$ onto the PLSDA model, a vector of predictions will be obtained ($\hat{y}_i$ is the maximum number of classes).

\[
\hat{y}_i = (\hat{y}_{i1}, \ldots, \hat{y}_{iJ})
\]

(7)

A simple approach is to assign the class membership to the maximum value over the $\hat{y}_i$ vector.

2.3. Prediction errors, confidence intervals and prediction intervals

Prediction estimates using either the jackknife or bootstrap were obtained by projecting unseen testing samples onto the calibration models obtained with these procedures (external validation procedure). Predictions were averaged using the root mean squared prediction error (RMSEP) [6].

A simple way to estimate confidence intervals is to sort the different estimates for a statistical parameter ($\theta$) and then to take the values corresponding to some percentile. These values or bounds will depend on the level of confidence required (e.g., 95%). This method has the advantage of being general and of being independent of the shape of the density function for that parameter. If the distribution is approximately normal, then the standard deviation ($s_\theta$) can be used to estimate the confidence limit for a certain probability level.

\[
\hat{\theta} = \bar{y} \pm t_{df, 1-\alpha/2} s_\theta
\]

(8)

The jackknife and bootstrap methods can be used here to estimate prediction intervals for the PLSDA predictions ($\hat{y}_i$ vector elements).

\[
\hat{y}_i = (\hat{y}_{i1} \pm t_{df, 1-\alpha/2} \text{SPE}_{ij}, \ldots, \hat{y}_{iJ} \pm t_{df, 1-\alpha/2} \text{SPE}_{ij})
\]

(9)

Prediction intervals for the SIMCA distances are estimated in the same manner as for the PLSDA predictions ($D_i$ vector elements).

\[
D_i = (D_{i1} \pm \text{SPE}_{ij}, \ldots, D_{iJ} \pm \text{SPE}_{ij})
\]

(10)

3. Experimental

3.1. Strains

A total of 105 isolates of multiresistant Acinetobacter baumannii, vancomycin-teicoplanin resistant Enterococcus faecium (VRE) and meticillin resistant Staphylococcus aureus/epidermidis (MRSA) from two Portuguese Hospitals during the years 2004 to 2007 were analyzed. Strains were isolated from feces, catheters, perianal swab, urine, respiratory samples, pus, blood and pleural fluid. Strains were tested for purity using nutrient agar medium. Plates were incubated at 37 °C for 24 h. Colonial growth was observed after incubation and isolated colonies were selected to evaluate by infrared spectroscopy. The samples set consisted of 49 isolates of Acinetobacter baumannii, 28 isolates of Enterococcus faecium, and 28 isolates of Staphylococcus aureus/epidermidis typed by PFGE (Pulse Field Gel Electrophoresis).

3.2. Molecular typing (golden method)

For the molecular typing PFGE was performed by a modification of a previously described method [28]. The strains were cultured in Nutrient Agar, 24 h at 37 °C and the DNA enclosed in agarose plugs. Restriction digestion of chromosomal DNA was performed with the enzymes SmaI (Enterococcus faecium and MRSA) and Apa I (Acinetobacter baumannii) overnight at 30 °C. The electrophoresis was performed in a “CHEF Mapper®XA Pulse Field Gel Electrophoresis System, BioRad”. The Lambda Ladder PFGE Marker (Biolabs) was used as a molecular size marker. The interpretation of the results was done by the bioinformatic program “BioNumerics, version 3.5” (Applied Maths BVBA, St-Martens-Latem, Belgium).

3.3. Phenotypic typing (FTIR method)

Strains were cultured in Tryptose soya agar (TSA), 24 h at 37 °C. Individually-selected colonies of each strain were carefully harvested from plates with a calibrated 1 µl loop. Colonies were suspended in 125 µl of sterile distilled water and homogenized in a vortex mixer. Thirty-five microliter aliquots of the bacterial suspensions were evenly applied onto each well in a silicium plate. Prior to analysis the samples were over dried at 44 °C for 40 min. Samples were run in triplicate and analyzed in a FT-IR spectroscopy using TENSOR spectrometer.
(Bruker Optik GmbH) in transmittance mode. Spectra were collected over the wavenumber range of 4000 cm$^{-1}$ to 600 cm$^{-1}$ with resolution of 4 cm$^{-1}$ under the control of a personal computer using "OPUS" software (Version 5.0, Bruker Optik GmbH). The OPUS spectral quality control test of the FT-IR spectra was applied routinely with thresholds for minimum absorbance (0.345) and maximum absorbance (1.245) for detector linearity, signal-to-noise (S/N) ratio, and water vapour.

3.4. Data analysis and pre-processing

All model calculations were carried out using Matlab version 6.5 release 13 (MathWorks, Natick, MA) and the PLS Toolbox version 3.5 for Matlab (Eigenvector Research, Manson, WA). The spectral pre-processing method was selected by taking into consideration the visual inspection of spectra PCA scores, explained variance, as well as uncertainty estimations. The best compromise was achieved by applying the multiplicative scatter correction and Savitzky–Golay filtering (first derivative, 9-points window size and second degree polynomial). The entire wavenumber region was considered in the analysis (4000–600 cm$^{-1}$). All spectra were mean-centred before models development.

4. Results and discussion

The raw and pre-processed FT-IR spectra of the available microbial samples were displayed in Fig. 1. For model interpretation purposes the 105 samples were first divided in a calibration and testing data set. The distribution of the samples between calibration and testing sets
was performed by randomly assigning each isolate to one of the sets in a way that the calibration set was composed by 2/3 of the total number of samples. The calibration data set was made by 73 samples and the testing data set retained the remaining 32 samples.

Spectra was initially analysed by PCA. A maximum number of 12 principal components were tested. The scores standard deviation for each testing sample obtained by bootstrapping the calibration data set was illustrated in Fig. 2. One can notice that for this particular data set, the scores standard deviation obtained after 1500 bootstraps (white circles) for the 32 validation samples. Standard deviation the entire validation data set is also plotted (grey circles).

**Fig. 2.** PCA scores standard deviation obtained after 1500 bootstraps (white circles) for the 32 validation samples. Standard deviation the entire validation data set is also plotted (grey circles).

**Fig. 3.** Score plot mapping the first against the second principal components of PCA with the application of bootstrapping. Calibration set scores are represented with black filled markers. Projected validation scores are represented with grey hollow markers (□ — *Acinetobacter baumannii*, ○ — *Staphylococcus aureus/Staphylococcus epidermidis*, Δ — *Enterococcus faecium*).
Fig. 4. $\chi^2$ distributions approximation to the histograms of the SIMCA distances using a 1 component model (a) for testing sample 1 (sample belonging to class 1); (b) for testing sample 23 (sample belonging to class 2); Class 1 — Acinetobacter baumanii, class 2 — Staphylococcus aureus/Staphylococcus epidermidis, class 3 — Enterococcus faecium.
the standard deviation decreases while the number of principal components increases only up to a certain point. Additionally, the standard deviation is very similar for every testing sample also only up to a certain number of principal components. The scores dispersion is almost identical for all samples up to four principal components, which indicates a good agreement between calibration and testing samples. It was observed a slight increase in the dispersion for the fifth, sixth and seventh principal components. It reflects the idea that at least for some samples the corresponding scores are no longer being accurately predicted by all bootstrapped PCA models. It is particularly evident that after the eighth principal component the PCA model is taking into account variations in the data that are no longer present in the testing samples (e.g., noise). In this case, it appears that four components can appropriately model both calibration and testing samples. The information related with uncertainty estimated in this way for information captured variance can provide a good estimate of the number of significant number of components.

The advantage of using a resampling method for samples classification with a multivariate non-supervised technique was analysed here. The evaluated technique was PCA, since this is the most common multivariate technique of microorganisms’ discrimination by FT-IR spectroscopy found in bibliography. Instead of having a single estimate for the scores (as in a usual projection), resampling methods allow a series of estimations to be obtained for each sample. By analysing the uncertainty of estimations it should be possible to define actual clusters. The calibration data set was bootstrapped and for each resampled data set a four component PCA model was obtained. Testing samples were projected onto each PCA model, yielding several estimates for each testing sample scores. In this case only 1500 bootstraps were performed. We can observe a clear distinction of groups representing Acinetobacter, Staphylococcus and Enterococcus isolates for the bootstrap strategy in Fig. 3. The discrimination level goes beyond the bacterial genus. It can be observed the formation of two clusters within the Staphylococcus group which correspond to Staphylococcus aureus and Staphylococcus epidermidis that are present in the data set. There was no visible overlapping between clouds of scores representing different classes which suggest in this case that no misclassification occurred. Although four components seemed to be more adequate to model the spectra, only two components were found to be enough for discrimination purposes. Jackknifing was also tested and results yielded similar result. For this reason, corresponding results were not shown. By plotting distinct estimations of scores it becomes clear the formation of clouds. These clouds define clearly if there are distinct clusters within the data set or not. This is very relevant in this context since discrimination is the target and the statistical significance of discriminant models is very important. Moreover, when only a few samples are available (there are situations were only a few isolates are available for a particular class) which is often the situation in these studies, this methodology appears to be more appropriate when compared to single calibration/testing results. Confidence intervals can be estimated from bootstraps for each testing sample score. If plotted in scatter plots like in Fig. 1, scores can be represented with corresponding bivariate distribution 95% confidence intervals.

The application and utility of uncertainty estimation in supervised multivariate models for classification purposes was also assessed. The two proposed supervised models were SIMCA and PLSDA. Bootstrap and jackknife were applied to the calibration data set 1500 times.

SIMCA models were developed to predict a class membership for new observations. For each new observation projected on the 3 PCA models (there are 3 classes), a vector of distances to the centre of each model was calculated as explained in the theory section. For each testing sample the class assignment corresponds to the minimum value over the three distances. In order to observe the advantage of being able to draw an empirical distribution for the distances observed for each sample an example is given for two testing samples: one representing Acinetobacter (testing sample 1), and other representing Staphylococcus (testing sample 23). A one component SIMCA model based on the entire calibration data set was tested with both testing samples. Both samples were correctly identified based on the criteria of minimum distance. However, this simple approach does not give any information about the precision of the estimates. Plots in Fig. 4a and b show $\chi^2$ distributions estimated with 1500 SIMCA bootstraps, approximated to the SIMCA distances for the two testing samples. The Acinetobacter sample can be easily identified as a member of class Acinetobacter, whether there are some doubts concerning the other sample due to existing overlaps between class Acinetobacter and class Staphylococcus distance distributions. With consideration to the calculated distance, sample was assigned to class Staphylococcus, which met the results obtained with molecular typing of mentioned sample. However, for the Staphylococcus sample, the statistical significance for the model assignment is unacceptable. It indicates that probably a one component SIMCA model is not enough. Mean distances and standard deviations obtained for all classes for different number of principal components are presented in Table 1. Both bias and standard deviation of test samples increase as the number of components increase. Ratios between mean distances and mean standard deviation values were calculated for each SIMCA model. The best number of components should encompass not only the distances information but also the standard deviation information. A graphical interpretation is provided in Fig. 5. It is clear that if a one component model was selected, then some discrimination problems are visible. The overlap for sample 23, already described, is visible in Fig. 5a. The analysis of the distances indicates that a two component model could predict the class membership for 100% of the test samples. The improvement is visible in Fig. 5b. One can notice that the standard deviation increases with the number of components. This suggests that keeping the SIMCA model with the lowest possible number of components is a better solution. Fig. 5c shows that the discrimination is not significantly better than the observed for two components (Fig. 5b) therefore we opted for the two components model.

PLS modelling was tested with the use of bootstrapping method as well. The 1500 bootstraps were generated for the calibration data set. A PLSDA model was calibrated from each data set and tested with the testing data set. Models up to 8 components were evaluated. The RMSEP and mean standard deviation obtained for the testing samples PLSDA predictions were compiled in Table 2. The analysis was repeated 100 times with a different selection of test set samples and

<table>
<thead>
<tr>
<th>Number of principal components</th>
<th>Class (average)</th>
<th>Mean distance</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA1</td>
<td>PCA2</td>
<td>PCA3</td>
<td>PCA1</td>
</tr>
<tr>
<td>1</td>
<td>0.708</td>
<td>1.663</td>
<td>3.599</td>
</tr>
<tr>
<td>2</td>
<td>1.815</td>
<td>0.786</td>
<td>3.679</td>
</tr>
<tr>
<td>3</td>
<td>2.100</td>
<td>1.973</td>
<td>0.873</td>
</tr>
<tr>
<td>4</td>
<td>0.827</td>
<td>2.219</td>
<td>4.404</td>
</tr>
<tr>
<td>5</td>
<td>2.649</td>
<td>0.864</td>
<td>4.605</td>
</tr>
<tr>
<td>6</td>
<td>3.181</td>
<td>2.994</td>
<td>0.973</td>
</tr>
<tr>
<td>7</td>
<td>0.899</td>
<td>2.903</td>
<td>5.570</td>
</tr>
<tr>
<td>8</td>
<td>3.425</td>
<td>0.945</td>
<td>4.888</td>
</tr>
<tr>
<td>9</td>
<td>4.453</td>
<td>4.622</td>
<td>0.873</td>
</tr>
<tr>
<td>10</td>
<td>1.036</td>
<td>3.468</td>
<td>6.803</td>
</tr>
<tr>
<td>11</td>
<td>4.528</td>
<td>1.001</td>
<td>7.226</td>
</tr>
<tr>
<td>12</td>
<td>6.022</td>
<td>5.174</td>
<td>1.158</td>
</tr>
<tr>
<td>13</td>
<td>1.306</td>
<td>4.068</td>
<td>8.020</td>
</tr>
<tr>
<td>14</td>
<td>6.167</td>
<td>1.090</td>
<td>8.469</td>
</tr>
<tr>
<td>15</td>
<td>8.299</td>
<td>6.310</td>
<td>1.254</td>
</tr>
<tr>
<td>16</td>
<td>1.599</td>
<td>4.878</td>
<td>10.172</td>
</tr>
<tr>
<td>17</td>
<td>8.367</td>
<td>1.196</td>
<td>10.823</td>
</tr>
<tr>
<td>18</td>
<td>11.468</td>
<td>7.774</td>
<td>1.486</td>
</tr>
</tbody>
</table>
Fig. 5. Testing samples and their SIMCA distances to the centre of PCA models. (a) one principal component, (b) two principal components and (c) three principal components (Acinetobacter baumanii (in black); Staphylococcus aureus/Staphylococcus epidermidis (in dark grey) and Enterococcus faecium (in light grey)).
our results show that obtained values do not depend much on the particular train and test data set. In our case, the RMSEP and standard deviation have a decreasing tendency as the number of latent variables increase. The greatest difference was observed from 1 to 2 latent variables. After the second latent variable, it can be observed a decreasing tendency for all classes but not as evident. If the bias information was the only one available, one might select two latent variables as the model dimension, since the bias decrease is not so important after that. This option is often based on the common assumption that the models robustness decrease with the number of latent variables. This situation was not observed here and the precision of estimations for 6 latent variables, for example, are better than those obtained for two components. A 6 latent variable PLSDA model was calibrated for further comparisons.

A comparison between jackknife and bootstrapping was performed for the supervised models using the previously identified best number of components (2 components for SIMCA and 6 components for PLSDA). The relevant results are marked in boldface in Table 3. Comparing the bootstrapping and jackknifing resampling strategies, the conclusion can only be that both yield similar results. No pattern could be identified for the bias or standard deviation estimations generated by the two methods. This was observed for the SIMCA model and PLSDA model. While comparing the results in terms of relative standard deviation for both model types, it can be observed that the precision in SIMCA is much lower than the precision with PLSDA. For example, dividing the mean standard deviation of *Acinetobacter* samples by the mean of predictions for PLSDA (relative standard deviation) yields $0.033/0.977*100=3.4\%$. The same calculation for SIMCA results in $0.209/0.818*100=25.5\%$. This difference is reflected in much wider confidence intervals for model estimations for SIMCA. This result was observed for the three different classes analysed, though. Therefore, the incorporation of the estimation of uncertainty in the modelling process demonstrated that PLSDA can generate more robust models (more precise estimations) than SIMCA. For this data set, both models were able to correctly identify 100% of the testing samples. These findings suggest that applying PLSDA to data sets where classes are harder to discriminate may be more adequate than SIMCA.

The described results were obtained for only one data set. Therefore, one could argue lack of universality and validity of the general findings. However, having analyzed other data sets (other microorganisms measured with FT-IR in transmittance mode with different number of samples) we reached the same general results. Therefore, we opted not to include the analysis of other similar data sets.

### Table 2

Bias and uncertainty for three different bacterial classes obtained with PLSDA and bootstrapping with a particular concern for number of latent variables used

<table>
<thead>
<tr>
<th>Number of latent variables</th>
<th>RMSEP</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Class 1</td>
<td>Class 2</td>
</tr>
<tr>
<td>1</td>
<td>0.422</td>
<td>0.236</td>
</tr>
<tr>
<td>2</td>
<td>0.176</td>
<td>0.143</td>
</tr>
<tr>
<td>3</td>
<td>0.128</td>
<td>0.111</td>
</tr>
<tr>
<td>4</td>
<td>0.085</td>
<td>0.079</td>
</tr>
<tr>
<td>5</td>
<td>0.058</td>
<td>0.072</td>
</tr>
<tr>
<td>6</td>
<td>0.051</td>
<td>0.060</td>
</tr>
<tr>
<td>7</td>
<td>0.041</td>
<td>0.051</td>
</tr>
<tr>
<td>8</td>
<td>0.038</td>
<td>0.044</td>
</tr>
</tbody>
</table>

5. Conclusions

The objective of this study was to assess the implementation of the jackknife and bootstrap strategies to evaluate the uncertainty of...
chemometrics methods in the context of discrimination and classification of microbial pathogens by FT-IR spectroscopy. Resampling strategies occurred to be very helpful as they allowed estimation of the optimal number of PCA components as well as to indicate the point where a PCA model takes into account possible variations that are no longer present in the testing data set. As an alternative for having a single estimate for the PCA scores, the jackknife and bootstrapping methods allow a cloud of estimations to be obtained for each sample and, therefore, enhancing the clusters definition for different kind of pathogens. This was observed for the analysed data set where clusters were positively identified.

Bootstrapping and jackknifing demonstrated to be two valid alternatives to estimate bias and variance for non-supervised and supervised models. SIMCA and PLSDA were the two supervised models tested. Both resampling techniques yielded similar results. Therefore, no particular advantage was detected for any technique. The knowledge about predictions variance proved to be useful to select the best number of model components. We showed that when comparing the estimated precision of PLSDA and SIMCA, the PLSDA predictions are more precise than the SIMCA predictions. The confidence intervals for PLSDA predictions are consequently narrower when compared to the SIMCA prediction intervals, indicating a superior robustness of the former model type.

Since the data set under study did not offer many difficulties concerning the separation of bacterial pathogens, a comparison between PLSDA and SIMCA would be required in terms of the application of resampling techniques at more difficult level (e.g. serotyping or phage typing) of microbial discrimination.

Acknowledgements

The authors wish to express their gratitude to Prof. J. Machado and Dr. R. Guimaraes (Enterobacteriaceae Unity, Bacteriology Center, National Health Institute Ricardo Jorge, Lisbon, Portugal) for kindly providing the data set that was used as the illustrative example in this study, as well as for valuable support and permission to publish the above results. O.P. gratefully acknowledges financial support from the Portuguese Foundation for Science and Technology (research grant no. SFRH/BD/15218/2004).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.chemolab.2008.06.005.