Accepted Manuscript

Uncertainty estimation and misclassification probability for classification models based on discriminant analysis and support vector machines

Camilo.L.M. Morais, Kássio.M.G. Lima, Francis.L. Martin

PII: S0003-2670(18)31091-2

DOI: 10.1016/j.aca.2018.09.022

Reference: ACA 236258

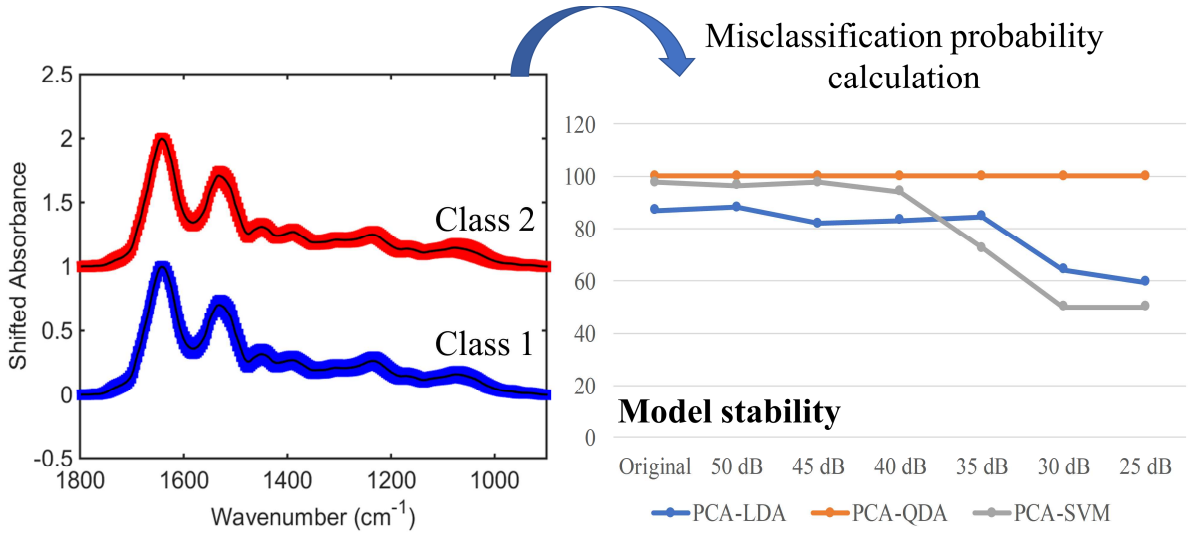
To appear in: *Analytica Chimica Acta*

Received Date: 19 July 2018

Revised Date: 5 September 2018

Accepted Date: 11 September 2018

Please cite this article as: C.L.M. Morais, K.M.G. Lima, F.L. Martin, Uncertainty estimation and misclassification probability for classification models based on discriminant analysis and support vector machines, *Analytica Chimica Acta* (2018), doi: https://doi.org/10.1016/j.aca.2018.09.022.



Uncertainty estimation and misclassification probability for classification models based on discriminant analysis and support vector machines

Camilo L. M. Morais[1]*, Kássio M. G. Lima[2], Francis L. Martin[1]

[1]School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, United Kingdom

[2]Biological Chemistry and Chemometrics, Institute of Chemistry, Federal University of Rio Grande do Norte, Natal 59072-970, Brazil

***Corresponding author:** Camilo L.M. Morais, Rm MB030, Maudland Building, School of Pharmacy and Biomedical Sciences, University of Central Lancashire, Preston PR1 2HE, United Kingdom; Tel.: +44 07476 704524; Email: cdlmedeiros-de-morai@uclan.ac.uk

Abstract

Uncertainty estimation provides a quantitative value of the predictive performance of a classification model based on its misclassification probability. Low misclassification probabilities are associated with a low degree of uncertainty, indicating high trustworthiness; while high misclassification probabilities are associated with a high degree of uncertainty, indicating a high susceptibility to generate incorrect classification. Herein, misclassification probability estimations based on uncertainty estimation by bootstrap were developed for classification models using discriminant analysis [linear discriminant analysis (LDA) and quadratic discriminant analysis (QDA)] and support vector machines (SVM). Principal component analysis (PCA) was used as variable reduction technique prior classification. Four spectral datasets were tested (1 simulated and 3 real applications) for binary and ternary classifications. Models with lower misclassification probabilities were more stable when the spectra were perturbed with white Gaussian noise, indicating better robustness. Thus, misclassification probability can be used as an additional figure of merit to assess model robustness, providing a reliable metric to evaluate the predictive performance of a classifier.

Keywords: Classification; Discriminant analysis; Figures of merit; Misclassification; Support vector machines; Uncertainty

1. Introduction

Multivariate classification models are commonly employed to segregate clusters based on a supervised learning approach. Commonly, the data are initially divided into training and external validation sets, where the first is used for model construction and the latter to assess the model performance. The predictive capacity of classification models is assessed by quality parameters also called “figures of merit”. The most used ones are the accuracy (total number of samples correct classified considering true and false negatives), sensitivity (proportion of positivies correctly identified) and specificity (proportion of negatives correctly identified) [1]. Additional figures of merit can also be estimated to confirm the predictive performance of a classification model, such as precision (classifier ability to avoid wrong predictions), F-score (overall performance of the model considering imbalanced data), G-score (overall performance of the model not accounting for class sizes), area under the curve (AUC) of receiver operating characteristic curves, positive and negative prediction values, positive and negative likelihood ratios, and Youden’s index [1-5]. The latter three are more commonly used for biomedical applications, where the ratio of true and false positives and negatives are an important factor towards making clinical decisions.

However, none of these figures of merit brings information of the degree of uncertainty in the classification model. Uncertainty is always present in any analytical measurement where a prior univariate or multivariate model is used to provide information of the property being analysed. For being non-specific, vibrational spectroscopy techniques generate thousands of data points for all chemical components that are susceptible to the radiation source incident on the sample, creating a very complex array of data for each sample analysed. To elucidate and extract information of the chemical components present in the spectrum, chemometric techniques are often employed. Multivariate calibration techniques, such as principal component regression (PCR) and partial least squares (PLS)

regression, are used for quantification applications; and classification techniques, such as discriminant analysis (DA) and support vector machines (SVM), for qualitative applications [6].

In spectroscopy applications, due to problems of collinearity and ill-conditioned data, variable reduction or selection techniques are often employed prior classification analysis. Principal component analysis (PCA) is one of the most popular methods of variable reduction, since it reduces all the spectral variables into a small number of principal components accounting for the majority of the original variance in the data [7]. Since the principal components are orthogonal to each other, the computation of inverse matrix operations used in discriminant analysis are achieved with high accuracy.

Uncertainty estimation for calibration models is well known [8, 9]. However, for classification techniques, uncertainty estimation is still a new topic, so far mainly explored for partial least squares discriminant analysis (PLS-DA) [10, 11]. Herein, we propose an uncertainty estimation method based on bootstrap for calculation of misclassification probabilities in linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and SVM models applied to four different datasets, where the classification stability is also evaluated by adding white Gaussian noise to the spectral data.

2. Experimental

2.1 Datasets

Four datasets were used for testing. Dataset 1 is composed of simulated spectra generated using a normal distribution function. Class 1 contains 30 spectra with 301 variables each, with mean ranging from 0.15 to 0.42 intensity units and standard deviation ranging from 0.41 to 1.14 intensity units between samples. Class 2 contains also 30 spectra with 301

variables each, with mean ranging from 0.19 to 0.35 intensity units and standard deviation ranging from 0.35 to 0.86 intensity units between sampels.

Dataset 2 is composed of 280 infrared (IR) spectra of two *Cryptococcus* fungi specimens acquired *via* attenuated total reflection Fourier-transform infrared (ATR-FTIR) spectroscopy. Class 1 contains 140 spectra of *Cryptococcus neoformans* samples, and class 2 contains 140 spectra of *Cryptococcus gattii* samples. Spectra were acquired in the range of 400-4000 cm^{-1} with resolution of 4 cm^{-1} and 16 co-added scans using a Bruker VEXTER 70 FTIR spectrometer (Bruker Optics Ltd., UK). The spectra were pre-processed by cut in the biofingerprint region (900-1800 cm^{-1}), followed by automatic weighted least squares baseline correction and normalisation to the Amide I peak (1650 cm^{-1}). More information about this dataset can be found in literature [12, 13].

Dataset 3 is composed of 240 IR spectra for two classes of formalin-fixed paraffin-embedded brain tissues measured using ATR-FTIR spectroscopy. Class 1 contains 140 spectra for normal brain tissue samples, and class 2 contains 100 spectra for glioblastoma brain tissue samples. Spectra were acquired in the range of 400-4000 cm^{-1} with resolution of 8 cm^{-1} and 32 co-added scans using a Bruker Vector 27 FTIR spectrometer with a Helios ATR attachement (Bruker Optics Ltd., UK). The spectra were pre-processed by cut in the biofingerprint region (900-1800 cm^{-1}), followed by ruberband baseline correction and normalisation to the Amide I peak (1650 cm^{-1}). This dataset is public available as part of IRootLab toolbox (http://trevisanj.github.io/irootlab/) [14, 15] and more information about it can be found in Gajjar et al. [16].

Dataset 4 is composed of 183 IR spectra separated into 3 classes. Class 1 is composed of 59 spectra of Syrian hamster embryo (SHE) cells contaminated with benzo[*a*]pyrene (B[*a*]P); class 2 is composed of 62 spectra of SHE cells contaminated with 3-

methylcholanthrene (3-MCA); and class 3 is composed of 62 spectra of SHE cells contaminated with anthracene (Ant). Spectra were acquired by using a Bruker TENSOR 27 spectrometer with a Helios ATR attachement (Bruker Optics Ltd., UK). Spectra were recorded in the range of 400-4000 cm^{-1} with a resolution of 8 cm^{-1}. Pre-processing was performed by cut in the biofingerprint region (900-1800 cm^{-1}), rubberband baseline correction and normalisation to the Amide I peak (1650 cm^{-1}). This dataset is public available as part of IRootLab toolbox (http://trevisanj.github.io/irootlab/) [14, 15]; further information can be found in Trevisan et al. [17].

2.2 Software

Data analysis was performed within MATLAB R2014b environment (The MathWorks, Inc., USA) using lab-made routines. Pre-processing was performed using PLS Toolbox 7.9.3 (Eigenvector Research, Inc., USA). Samples were divided into training (70%) and external validation (30%) sets using Kennard-Stone sample selection algorithm [18].

2.3 Classification techniques

Data were initially processed by PCA in order to reduce the number of variables and solve ill-condition problems. PCA decomposes the original spectral matrix **X** into scores (**T**), loadings (**P**) and residuals (**E**) as follows [7]:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \quad (1)$$

The PCA scores were used as input variables for the classification models (LDA, QDA and SVM) with the number of principal components selected by singular value decomposition (SVD) [1, 7] and root mean square error of cross-validation (RMSECV) values obtained with cross-validated PCA [19]. The cumulated explained variance was calculated based on SVD as follows [1]:

$$\mathbf{X} = \mathbf{USV}^{-1} \tag{2}$$

$$\mathrm{v}(\%) = \left[\frac{\mathrm{diag}(\mathbf{S})}{\sum \mathrm{diag}(\mathbf{S})}\right] \times 100 \tag{3}$$

where v(%) is the explained variance; **U** and **V** are orthogonal matrices; and **S** is a matrix containing nonzero singular values on its diagonal.

The LDA (L_{ik}) and QDA (Q_{ik}) classification scores were calculated in a non-Bayesian form as follows [20, 21]:

$$L_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \mathbf{C}_{\mathrm{pooled}}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \tag{4}$$

$$Q_{ik} = (\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \mathbf{C}_k^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_k) \tag{5}$$

where $\mathbf{x}_i$ are the input variables for sample i; $\bar{\mathbf{x}}_k$ is the mean vector of class k; $\mathbf{C}_{\mathrm{pooled}}$ is the pooled covariance matrix; and $\mathbf{C}_k$ is the variance-covariance matrix of class k. $\mathbf{C}_k$ and $\mathbf{C}_{\mathrm{pooled}}$ are estimated as follows:

$$\mathbf{C}_k = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^{\mathrm{T}} \tag{6}$$

$$\mathbf{C}_{\mathrm{pooled}} = \frac{1}{n} \sum_{k=1}^{K} n_k \mathbf{C}_k \tag{7}$$

where n_k is the number of samples of class k; n is the total number of samples in the training set; and K is the number of classes.

SVM was applied to the PCA scores using a radial basis function (RBF) kernel [22]. The SVM classifier takes the form of [13]:

$$r_i = \mathrm{sign}\left(\sum_{i=1}^{N_{SV}} \alpha_i y_i K(\mathbf{x}_i, \mathbf{z}_j) + b\right) \tag{8}$$

where r_i is the classification response for sample i; N_{SV} is the number of support vectors; α_i is the Lagrange multiplier; y_i is the class membership (±1) of sample i; $K(\mathbf{x}_i, \mathbf{z}_j)$ is the kernel function; and b is the bias parameter.

2.4 Misclassification probability estimation

The uncertainty estimation was based on Bootstrap [23], a random sampling method with replacement that allows confidence intervals to be placed on the model predictions based on uncertainties of the original data [10]. The procedure for calculating uncertainties based on residual bootstrap was originally presented by de Almeida et al. [11] and adapted herein for LDA, QDA and SVM-based models. For comparison, uncertainty propagation estimate for SVM was calculated by differentiation of Eq. 8 based on a previous uncertainty estimation for RBF kernel in artificial neural networks (ANN), assuming that noise only affects the test sample [24]:

$$\mathrm{d}r = \sum_{i=1}^{N_{SV}} \alpha_i y_i \frac{\mathrm{d}K(\mathrm{x}_i, \mathrm{z}_j)}{\mathrm{d}x_i} \mathrm{d}x_i = \mathbf{b}_{\mathrm{SVM}}^{\mathrm{T}} \mathrm{d}\mathbf{x} \quad (9)$$

where $\mathbf{b}_{\mathrm{SVM}}^{\mathrm{T}}$ represents the uncertainty propagation of SVM using RBF kernel.

For bootstrap uncertainty estimation, initially, the residuals for LDA, QDA or SVM models are calculated using:

$$\mathbf{f}^* = \frac{\mathbf{f}}{\sqrt{1 - {}^{D_f}/_{n}}} \quad (10)$$

where $\mathbf{f}^*$ is the weighted model residual; $\mathbf{f}$ is the model residual; and D_f is the pseudo-degrees of freedom [25]. $\mathbf{f}$ is estimated for LDA, QDA or SVM models as:

$$\mathbf{f} = \mathbf{y} - \hat{\mathbf{y}} \quad (11)$$

where $\mathbf{y}$ is the reference class category for all samples; and $\hat{\mathbf{y}}$ is the model response for LDA [$\hat{\mathbf{y}} = (L_1, \cdots, L_n)$]; QDA [$\hat{\mathbf{y}} = (Q_1, \cdots, Q_n)$]; or SVM [$\hat{\mathbf{y}} = (r_1, \cdots, r_n)$].

Then, bootstrapping is applied by removing sample i whose uncertainty is being estimated by the model. A new response matrix $\mathbf{y}^*$ is generated by replacing the remaining values in $\mathbf{y}$ with the model predicted $\hat{\mathbf{y}}$. Then, a new random residual vector $\mathbf{f}^*_{\text{boot}}$ is generated by bootstrapping. The bootstrapping residual $\mathbf{f}^*_{\text{boot}}$ is added to the $\hat{\mathbf{y}}$ predicted, generating a new response vector $\mathbf{y}^{**}$:

$$\mathbf{y}^{**} = \hat{\mathbf{y}} + \mathbf{f}^*_{\text{boot}} \tag{12}$$

A new classification model is then created using $\mathbf{y}^{**}$ as reference categories. Finally, a new residual vector $\hat{\mathbf{f}}^*$ is created by subtracting the bootstrapping predicted values $\hat{\mathbf{y}}^{**}$ from the model predicted $\hat{\mathbf{y}}$:

$$\hat{\mathbf{f}}^* = \hat{\mathbf{y}} - \hat{\mathbf{y}}^{**} \tag{13}$$

The confidence intervals are calculated for sample i based on the residual vector $\hat{\mathbf{f}}^*$. For a 95% confidence interval, the lower bound ($\mathbf{c}_{\text{low}}$) and the upper bond ($\mathbf{c}_{\text{up}}$) are given by:

$$\mathbf{c}_{\text{low}} = 0.25\hat{\mathbf{f}}^* \tag{14}$$

$$\mathbf{c}_{\text{up}} = 0.975\hat{\mathbf{f}}^* \tag{15}$$

For misclassification probability calculation, the classification categories $\mathbf{y}$ are treated as being normally distributed with mean equal to $\hat{\mathbf{y}}$ and standard deviation $\sigma = {}^1\!/_4\left(\mathbf{c}_{\text{low}} - \mathbf{c}_{\text{up}}\right)$. The probability that sample i is class k=1, denoted $P_{1,i}$, is equivalent to the probability that $\hat{y}_i$ is lower than the threshold value that separates the classes, y_{bound}. $P_{1,i}$ is given by the cumulative distribution function for the normal distribution [10]:

$$P_{1,i} = P(\hat{y}_i \leq y_{\text{bound}}) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{y_{\text{bound}} - \hat{y}_i}{\sqrt{2}\sigma_i}\right)\right] \quad (16)$$

Similarly, the probability that sample i is class k=2, denoted $P_{2,i}$, is equal to $1 - P_{1,i}$. The misclassification probability of sample i, $m_{p,i}$, is therefore determined based on the classification of sample i as:

$$m_{p,i} = P_{1-y_i} \quad (17)$$

The m_p values range from 0 (no misclassification probability) to 1 (maximum misclassification probability). Values above 0.5 indicate higher probability of misclassification. A graphical flowchart illustrating the processing steps for misclassification probability calculation for PCA-LDA, PCA-QDA and PCA-SVM models is depicted in Figure 1.

[Insert Figure 1 here]

3. Results and discussion

Datasets 1-4 were analysed in order to estimate the misclassification probability associated with the trustworthiness and robustness of three classification algorithms: PCA-LDA, PCA-QDA and PCA-SVM. Pre-processed spectra with mean and standard-deviation for these datasets are depicted in Figure 2.

[Insert Figure 2 here]

Dataset 1 is composed of simulated spectra (Figure 2a). Although this dataset has no chemical meaning, simulated data are commonly used as a primary source to evaluate discriminatory performance of classification algorithms [1]. PCA was applied to the data and 10 PCs were selected according to SVD and RMSECV values (Figure 3a and b) (cumulative

variance of 97.2%). PCA-LDA did not show a good classification, with an accuracy of 44.4%. The average misclassification rate for the test set was equal to 0.520. This high misclassification probability indicates a large degree of uncertainty for the PCA-LDA model, which is confirmed by the high misclassification probability (m_p > 0.5). On the other hand, by applying a QDA classifier, the classification performance improved substantially. The accuracy in the external validation set was found at 88.9% with average misclassification probability of 0.113. QDA performance was superior than the one found by LDA due to the difference variance structures of class 1 and 2, as evidenced in the standard-deviation in Figure 2a. LDA assumes classes having similar variance structures, using a pooled covariance model. In contrast, QDA assumes classes having different variance structures, which improves considerably its performance over LDA when this condition happens [20, 21]. Additional figures of merit are depicted in Table 1.

[Insert Figure 3 here]

[Insert Table 1 here]

SVM was applied to the PCA scores by means of PCA-SVM generating also a good prediction response (accuracy = 94.4%). Although SVM fitting and prediction are better than QDA in terms of accuracy, sensitivity and specificity; its average misclassification probability is slightly higher (m_p = 0.152). A robustness test was then performed by adding white Gaussian noise to the spectra in 6 different levels of signal-to-noise ratio (S/N) measured in decibels (dB). S/N values of 50 dB, 45 dB, 40 dB, 35 dB, 30 dB and 25 dB were tested. As can be seen in Figure 4a, by adding noise to the spectra, the predictive performance in terms of overall accuracy remained constant for PCA-QDA and PCA-SVM models. For PCA-LDA, the addition of noise at 25 dB improved the accuracy to 50%. This phenomenon could happen due to the poor-fitting of the LDA model for dataset 1 (sensitivity and

specificity of 44.4%), since in this case the model response might not be entirely reliable on the signal quality.

[Insert Figure 4 here]

For dataset 2 (*Cryptococcus* fungi specimens), PCA-QDA also had a better performance than PCA-LDA. According to Figure 2b, class 1 has a clear higher variance for the variables in the range of 900-1200 cm^{-1} (phosphodiester, polysaccharides, glycogen and PO_2^- symmetric stretching in DNA/RNA [26]) in comparison with class 2. PCA-QDA achieved perfect class segregation (accuracy = 100%), while PCA-LDA achieved fair results with accuracy at 86.9%. All models were built using 8 PCs determined by SVD and RMSECV values (Figure 3c and d) (cumulative variance of 99.8%). Average misclassification probabilities of 0.328 and 0.212 were found for LDA and QDA models, respectively; confirming the higher trustworthiness of PCA-QDA over PCA-LDA for this dataset (Table 1). PCA-SVM also achieved good classification results, with an accuracy of 97.6% in the external validation set. However, the average misclassification probability was found at 0.500, which indicates that this model is not stable. The negative predictive value (NPV) for PCA-SVM indicates that the presence of misclassification is present only in the negative samples (*Cryptococcus neoformans*), a possible overfitting sign. Robustness was again evaluated by adding white Gaussian noise to the spectra set. The PCA-QDA was the only model that remained stable with noise, while the other two models (PCA-LDA and PCA-SVM) had an accentuated decrement of accuracy after S/N of 40 dB (Figure 4b). As expected by the misclassification probabilities values, the performance of PCA-SVM when the spectra were perturbated by noise was even worse than using PCA-LDA, since its accuracy dropped to 50% at 25 dB.

Dataset 3 is composed of IR spectra of normal brain tissue samples (class 1) and glioblastoma brain tissue samples (class 2) (Figure 2c). Both classes seem to have similar spectral profiles and standard-deviations. PCA-SVM classified the data with 100% accuracy (misclassification probability of 0.244) using 10 PCs selected by SVD and RMSECV values (Figure 3e and f) (cumulative variance of 99.4%). The second best classification performance was found using PCA-QDA (accuracy = 88.9%, misclassification probability of 0.276) and, for last, PCA-LDA (accuracy = 68.1, misclassification probability of 0.319). The three models are stable until S/N 35 dB, but after this point, all the classifiers tend to lose their classification performance converging to accuracies of 54.2% (PCA-LDA), 58.3% (PCA-QDA) and 62.5% (PCA-SVM) at 25 dB (Figure 3c).

Dataset 4 is composed of 3 classes of samples measured by ATR-FTIR. The average spectra with standard-deviation for class 1 (SHE cells contaminated with B[*a*]P), class 2 (SHE cells contaminated with 3-MCA) and class 3 (SHE cells contaminated with Ant) are depicted in Figure 2d. The variance among the classes seem to be evenly distributed, according to the similar standard-deviation observed in Figure 2d. PCA-LDA was applied using 10 PCs selected by SVD and RMSECV values (Figure 3g and h) (cumulative variance of 98.9%), generating an overall accuracy of 91.1% (average misclassification probability = 0.260). This model had the best classification performance in comparison with PCA-QDA and PCA-SVM, which seem to be overfitted according to the small sensitivity and specificity values observed between the classes (Table 1). PCA-QDA achieved an overall accuracy of 75.0% (average misclassification probability of 0.384) and PCA-SVM with an overall accuracy of 90.4% (average misclassification probability of 0.406). By applying noise to the data (Figure 4d), the model performance for PCA-LDA remained constant until 35 dB, then quickly dropped afterwards. For PCA-SVM, the model maintained overall accuracy around 90% until 40 dB, followed by a quickly dropping at 35 dB; and for PCA-QDA, the overall

accuracy decreased steadily until 25 dB. At 25 dB, all models converged to the same accuracy of 57%.

The mean misclassification probability and uncertainty propagation estimate based on Eq. 9 for SVM models are compared in Figure 4. An exponential trend is observed between the two parameters (Figure 4a), where the uncertainty propagation is proportional to the misclassification probability. A linear relationship between the two parameters is depicted in Figure 4b by the application of a natural logarithm function, where an R^2 of 0.971 is found; indicating that the classification uncertainty by bootstrap behaves similar to that one found using RBF functions [24].

[Insert Figure 5 here]

4. Conclusion

Misclassification probabilities were determined for PCA-LDA, PCA-QDA and PCA-SVM models applied to 4 different datasets (1 simulated and 4 real data). Uncertainty estimations were calculated by bootstrapping in order to obtain confidence intervals for misclassification probability calculations, presented herein as a new quality parameter to indicate model trustworthiness for these three classifiers. A correlation between the misclassification probability and model robustness was observed by adding white Gaussian noise to the spectral datasets, in which models with higher misclassification probabilities were more susceptible to error. Therefore, the misclassification probability can be used as a new figure of merit to assess model quality in classification applications, containing information of the model uncertainty and being also used to evaluate model robustness.

Acknowledgments

Camilo L. M. Morais would like to thank CAPES-Brazil (grant 88881.128982/2016-01) for financial support.

References

[1] C.L.M. Morais, K.M.G. Lima, Comparing unfolded and two-dimensional discriminant analysis and support vector machines for classification of EEM data, Chemometr. Intell. Lab. Syst. 170 (2017) 1-12.

[2] K.S. Parikh, T.P. Shah, Support Vector Machine – a Large Margin Classifier to Diagnose Skin Illnesses, Procedia Technol. 23 (2016) 369-375.

[3] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemometr. Intell. Lab. Syst. 174 (2018) 33-44.

[4] A.C.O. Neves, C.L.M. Morais, T.P.P. Mendes, B.G. Vaz, K.M.G. Lima, Mass spectrometry and multivariate analysis to classify cervical intraepithelial neoplasia from blood plasma: an untargeted lipidomic study, Sci. Rep. 8 (2018) 3954.

[5] L.F.S. Siqueira, R.F. Araújo Júnior, A.A. de Araújo, C.L.M. Morais, K.M.G. Lima, LDA vs. QDA for FT-MIR prostate cancer tissue classification, Chemometr. Intell. Lab. Syst. 162 (2017) 123-129.

[6] T. Naes, T. Isaksson, T. Fearn, T. Davies, A User-Friendly Guide to Multivariate Calibration and Classification, NIR Publications, Chichester, 2002.

[7] R. Bro, A.K. Smilde, Principal component analysis, Anal. Methods 6 (2014) 2812-2831.

[8] D.G. Cacuci, M. Ionescu-Bujor, Sensitivity and Uncertainty Analysis, Data Assimilation, and Predictive Best-Estimate Model Calibration, in: D.G. Cacuci (Eds.), Handbook of Nuclear Engineering, Springer, Boston, 2010, pp. 1913-2051.

[9] J. Caja, E. Gómez, P. Maresca, Optical measuring equipments. Part I: Calibration model anduncertainty estimation, Precision Engineering 40 (2015) 298-304.

[10] W.F.C. Rocha, D.A. Sheen, Classification of biodegradable materials using QSAR modelling with uncertainty estimation, SAR QSAR Environ. Res. 27 (2016) 799-811.

[11] M.R. de Almeida, D.N. Correa, W.F.C. Rocha, F.J.O. Scafi, R.J. Poppi, Discrimination between authentic and counterfeit banknotes using Raman spectroscopy and PLS-DA with uncertainty estimation, Microchem. J. 109 (2013) 170-177.

[12] F.S.L. Costa, P.P. Silva, C.L.M. Morais, T.D. Arantes, E.P. Milan, R.C. Theodoro, K.M.G. Lima, Attenuated total reflection Fourier transforminfrared (ATR-FTIR) spectroscopy as a new technology for discrimination between *Cryptococcus neoformans* and *Cryptococcus gattii*, Anal. Methods 8 (2016) 7107-7115.

[13] C.L.M. Morais, F.S.L. Costa, K.M.G. Lima, Variable selection with a support vector machine for discriminating *Cryptococcus* fungal species based on ATR-FTIR spectroscopy, Anal. Methods 9 (2017) 2964-2970.

[14] J. Trevisan, P.P. Angelov, A.D. Scott, P.L. Carmichael, F.L. Martin, IRootLab: a free and open-source MATLAB toolbox for vibrational biospectroscopy data analysis, Bioinformatics 29 (2013) 1095-1097.

[15] F.L. Martin, J.G. Kelly, V. Llabjani, P.L. Martin-Hirsch, I.I. Patel, J. Trevisan, N.J. Fullwood, M.J. Walsh, Distinguishing cell types or populations based on the computational analysis of their infrared spectra, Nat. Protoc. 5 (2010) 1748-1760.

[16] K. Gajjar, L.D. Heppenstall, W. Pang, K.M. Ashton, J. Trevisan, I.I. Patel, V. Llabjani, H.F. Stringfellow, P.L. Martin-Hirsch, T. Dawson, F.L. Martin, Diagnostic segregation of human brain tumours using Fourier-transform infrared and/or Raman spectroscopy coupled with discriminant analysis, Anal. Methods 5 (2013) 89-102.

[17] J. Trevisan, P.P. Angelov, I.I. Patel, G.M. Najand, K.T. Cheung, V. Llabjani, H.M. Pollock, S.W. Bruce, K. Pant, P.L. Carmichael, A.D. Scott, F.L. Martin, Syrian hamster embryo (SHE) assay (pH 6.7) coupled with infrared spectroscopy and chemometrics towards toxicological assessment, Analyst 135 (2010) 3266-3272.

[18] R.W. Kennard, L.A. Stone, Computer Aided Design of Experiments, Technometrics 11 (1969) 137-148.

[19] R.G. Brereton, Chemometrics. Data Analysis for the Laboratory and Chemical Plant, Wiley, Chichester, 2003.

[20] S.J. Dixon, R.G. Brereton, Comparison of performance of five common classifiers represented as boundary methods: Euclidean Distance to Centroids, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Learning Vector Quantization and Support Vector Machines, as dependent on data structure, Chemometr. Intell. Lab. Syst. 95 (2009) 1-17.

[21] C.L.M. Morais, K.M.G. Lima, Principal Component Analysis with Linear and Quadratic Discriminant Analysis for Identification of Cancer Samples Based on Mass Spectrometry, J. Braz. Chem. Soc. 29 (2018) 472-481.

[22] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273-297.

[23] R. Wehrens, H. Putter, L.M.C. Buydens, The bootstrap: a tutorial, Chemometr. Intell. Lab. Syst. 54 (2000) 35-52.

[24] F. Allegrini, A.C. Olivieri, Sensitivity, Prediction Uncertainty, and Detection Limit for Artificial Neural Network Calibrations, Anal. Chem. 88 (2016) 7807-7812.

[25] H. van der Voet, Pseudo-degrees of freedom for complex predictive models: the example of partial least squares, J. Chemometrics 13 (1999) 195-208.

[26] Z. Movasaghi, S. Rehman, I. ur Rehman, Fourier Transform Infrared (FTIR) Spectroscopy of Biological Tissues, Appl. Spectrosc. Rev. 43 (2008) 134-179.

Captions for Figures

Figure 1: Flowchart illustrating data processing steps for misclassification probability calculation. D_f stands for pseudo-degrees of freedom.

Figure 2: Mean and standard-deviation (shaded area) for (a) dataset 1, (b) dataset 2, (c) dataset 3, and (d) dataset 4.

Figure 3: Singular value decomposition (SVD) for (a) dataset 1, (c) dataset 2, (e) dataset 3 and (g) dataset4; root mean square error of cross-validation (RMSECV) of PCA for (b) dataset 1, (d) dataset 2, (f) dataset 3 and (h) dataset 4 varying the number of principal components (PCs).

Figure 4: Overall accuracy in percentage for PCA-LDA, PCA-QDA and PCA-SVM models in (a) dataset 1, (b) dataset 2, (c) dataset 3 and (d) dataset 4, by adding white Gaussian noise to the spectra datasets in the following levels of signal-to-noise ratio: 50 dB, 45 dB, 40 dB, 35 dB, 30 dB and 25 dB.

Figure 5: (a) Mean misclassification probability using bootstrap *versus* norm of uncertainty propagation coefficients ($\mathbf{b}_{\mathrm{SVM}}^{\mathrm{T}}$) calculated for SVM models with the training samples of datasets 1–4; and **(b)** mean misclassification probability using bootstrap *versus* natural logarithm of the norm of uncertainty propagation coefficients ($\mathbf{b}_{\mathrm{SVM}}^{\mathrm{T}}$) calculated for SVM models with the training samples of datasets 1–4 (linear equation: $y = 13.3x + 1.13$).

Table 1: Figures of merit calculated for the external validation set in datasets 1–4. PPV stands for positive predictive value, NPV for negative predictive value, YOU for Youden's index, and m_p stands for average misclassification probability.

Dataset 1	Accuracy	Sensitivity	Specificity	PPV	NPV	YOU	m_p
PCA-LDA	44.4%	44.4%	44.4%	44.4%	44.4%	-11.1%	0.520
PCA-QDA	88.9%	77.8%	100%	100%	81.8%	77.8%	0.113
PCA-SVM	94.4%	88.9%	100%	100%	90.0%	88.9%	0.152
Dataset 2							
PCA-LDA	86.9%	97.6%	76.2%	80.4%	97.0%	73.8%	0.328
PCA-QDA	100%	100%	100%	100%	100%	100%	0.212
PCA-SVM	97.6%	95.2%	100%	100%	95.5%	95.2%	0.500
Dataset 3							
PCA-LDA	68.1%	80.0%	59.5%	58.5%	80.6%	39.5%	0.319
PCA-QDA	88.9%	90.0%	88.1%	84.4%	92.5%	78.1%	0.276
PCA-SVM	100%	100%	100%	100%	100%	100%	0.244
Dataset 4							
PCA-LDA							
Class 1	94.6%	94.7%	94.4%	97.3%	89.5%	89.2%	0.265
Class 2	89.3%	83.8%	100%	100%	76.0%	83.8%	0.217
Class 3	89.3%	91.9%	84.2%	91.9%	84.2%	76.1%	0.299
PCA-QDA							
Class 1	76.8%	100%	27.8%	74.5%	100%	27.8%	0.500
Class 2	73.2%	100%	21.1%	71.2%	100%	21.1%	0.434
Class 3	75.0%	62.2%	100%	100%	57.6%	62.2%	0.217
PCA-SVM							
Class 1	98.2%	97.4%	100%	100%	94.7%	97.4%	0.447
Class 2	100%	100%	100%	100%	100%	100%	0.468
Class 3	73.2%	59.5%	100%	100%	55.9%	59.5%	0.303

Figure 1

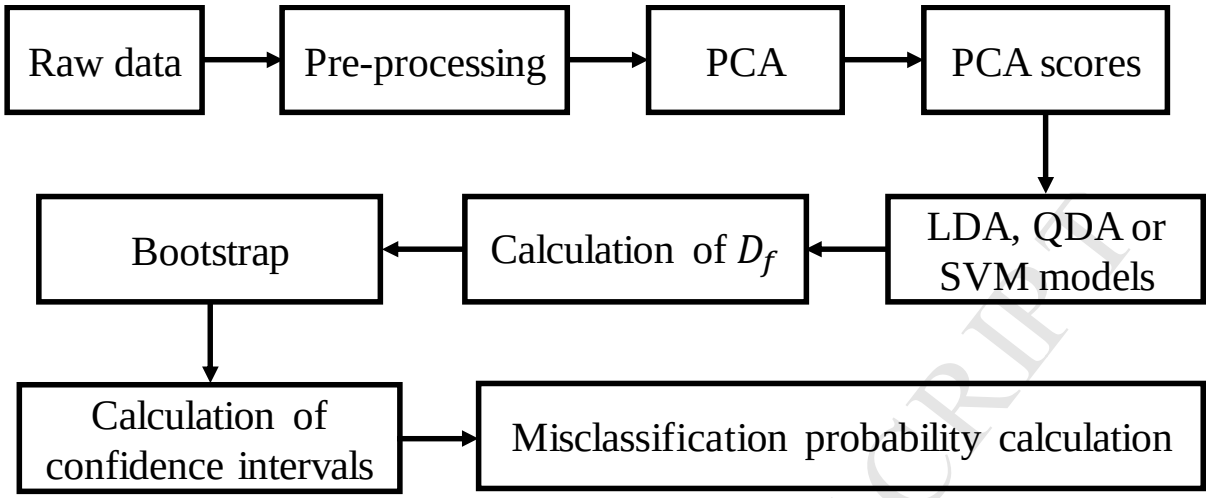


Figure 2

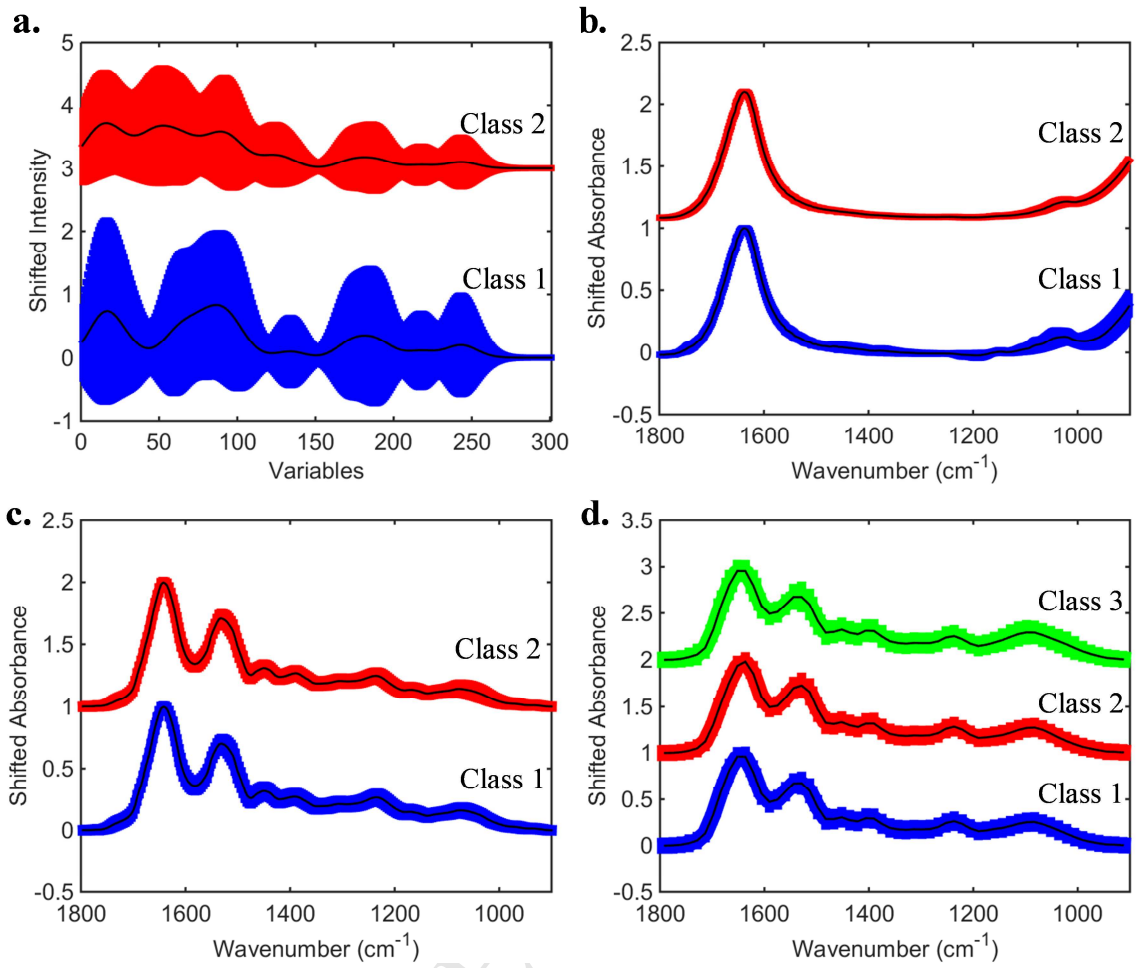


Figure 3

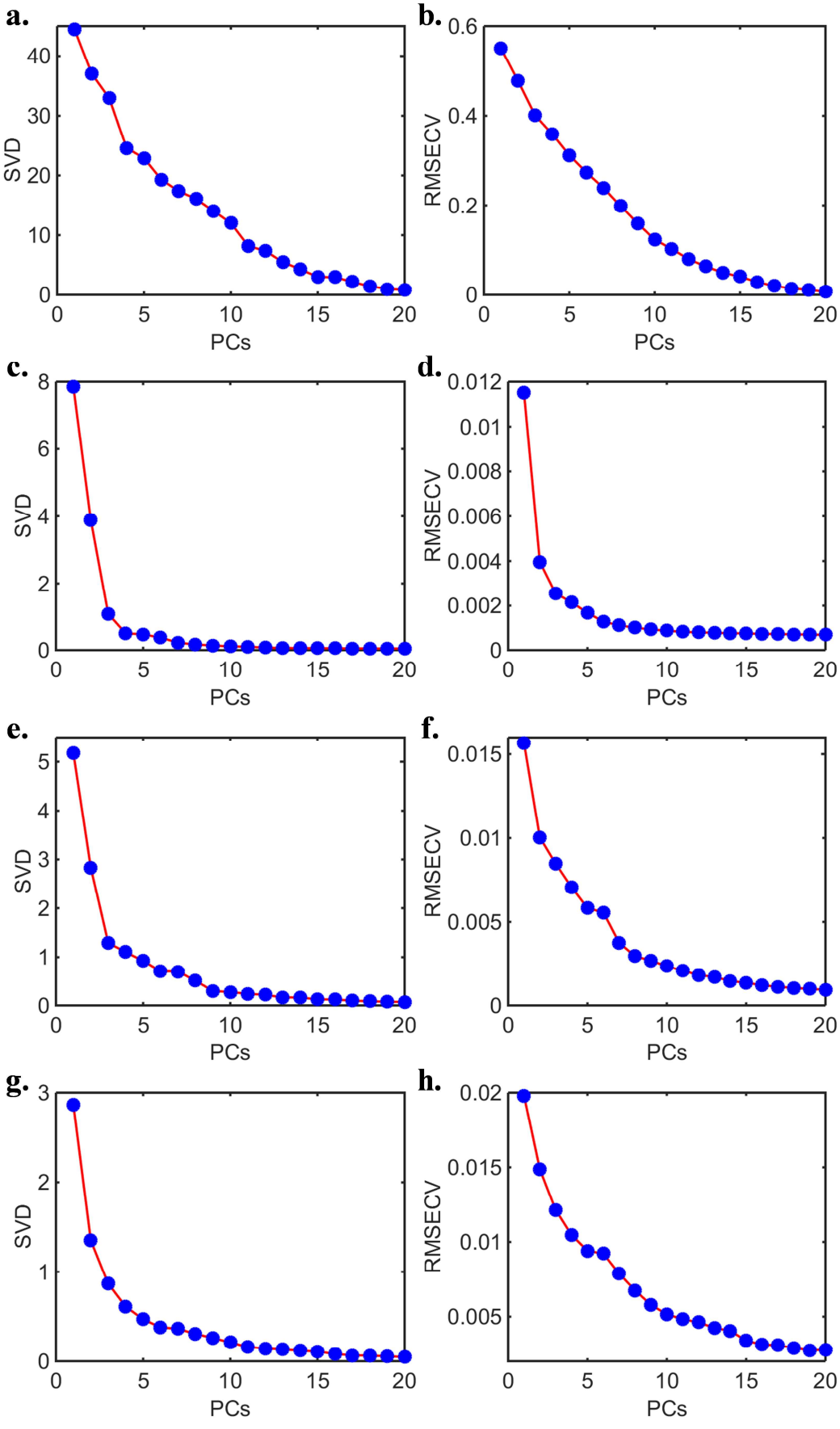

a.
SVD
PCs
b.
RMSECV
PCs
c.
SVD
PCs
d.
RMSECV
PCs
e.
SVD
PCs
f.
RMSECV
PCs
g.
SVD
PCs
h.
RMSECV
PCs

Figure 4

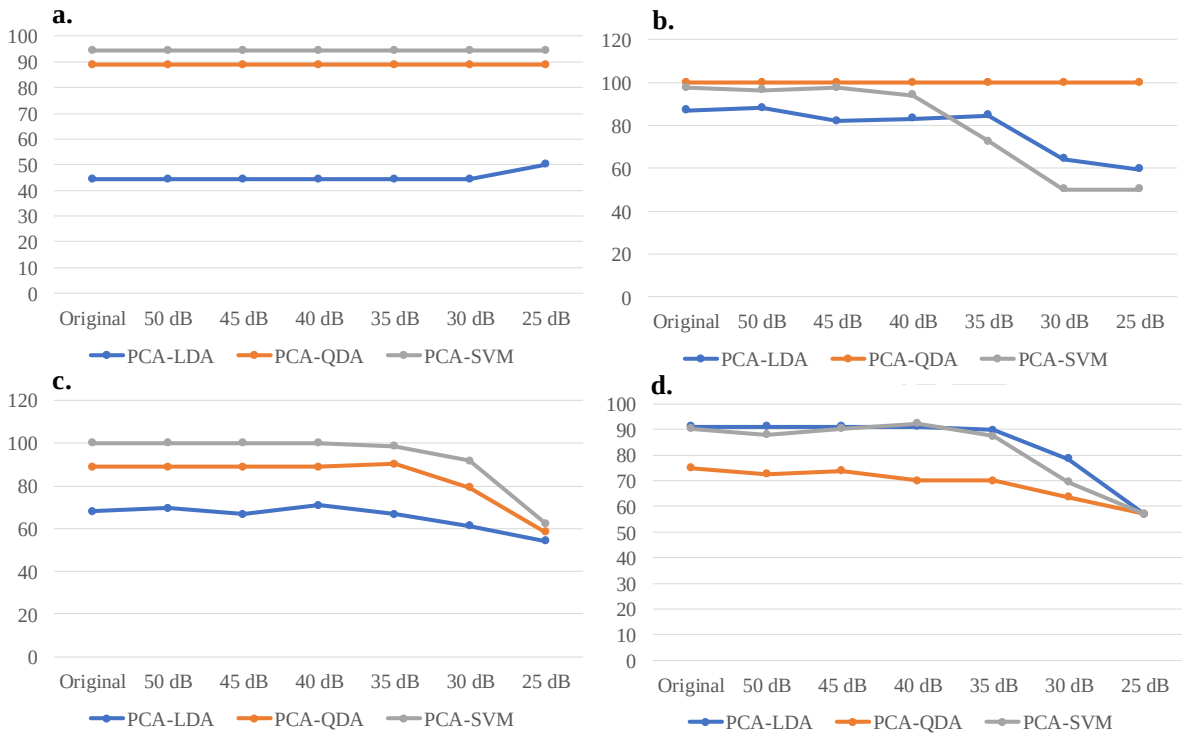
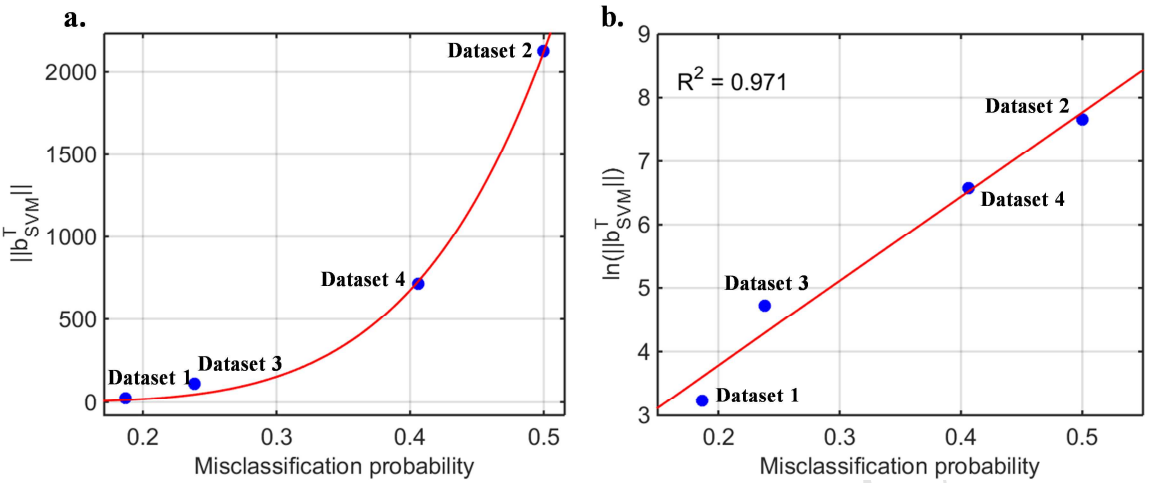
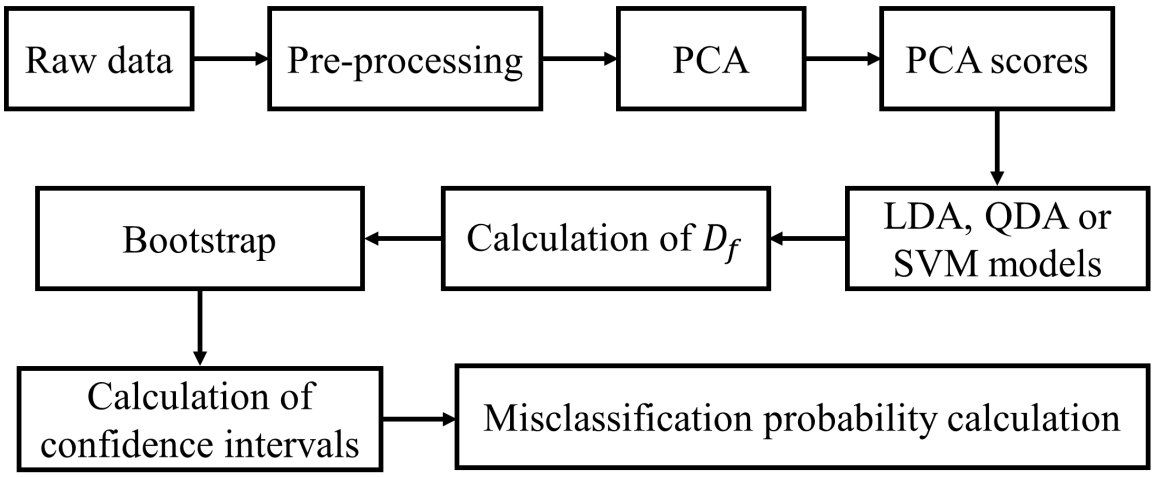
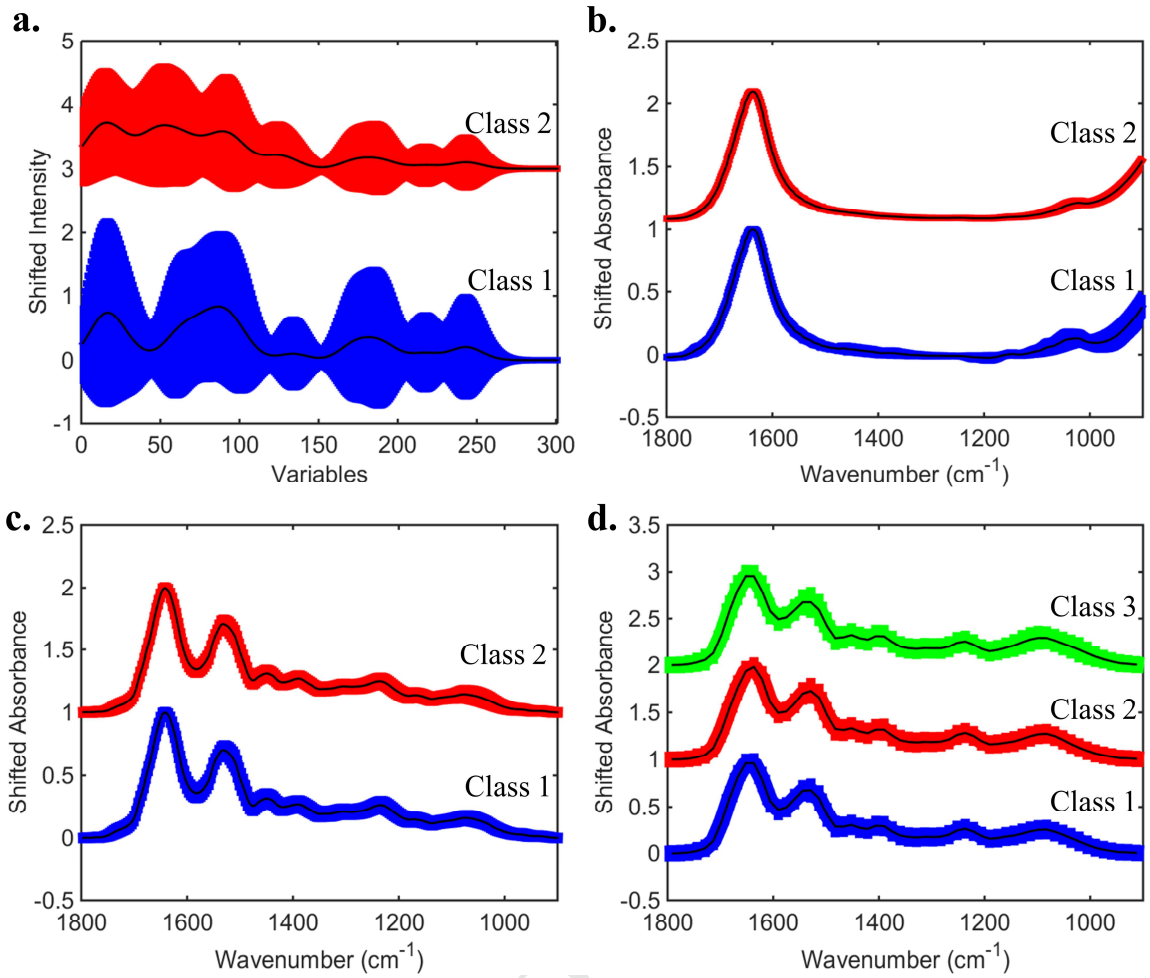
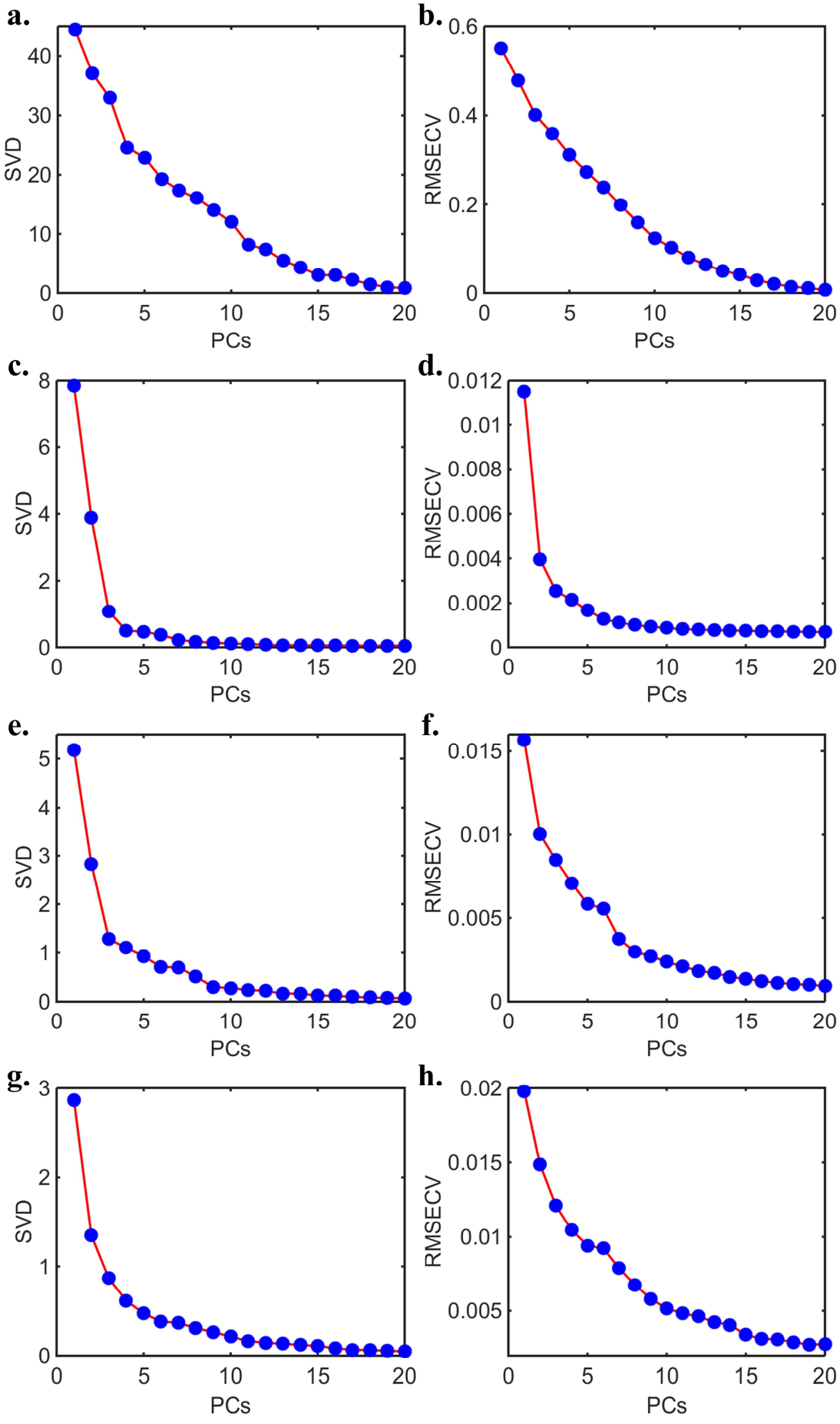


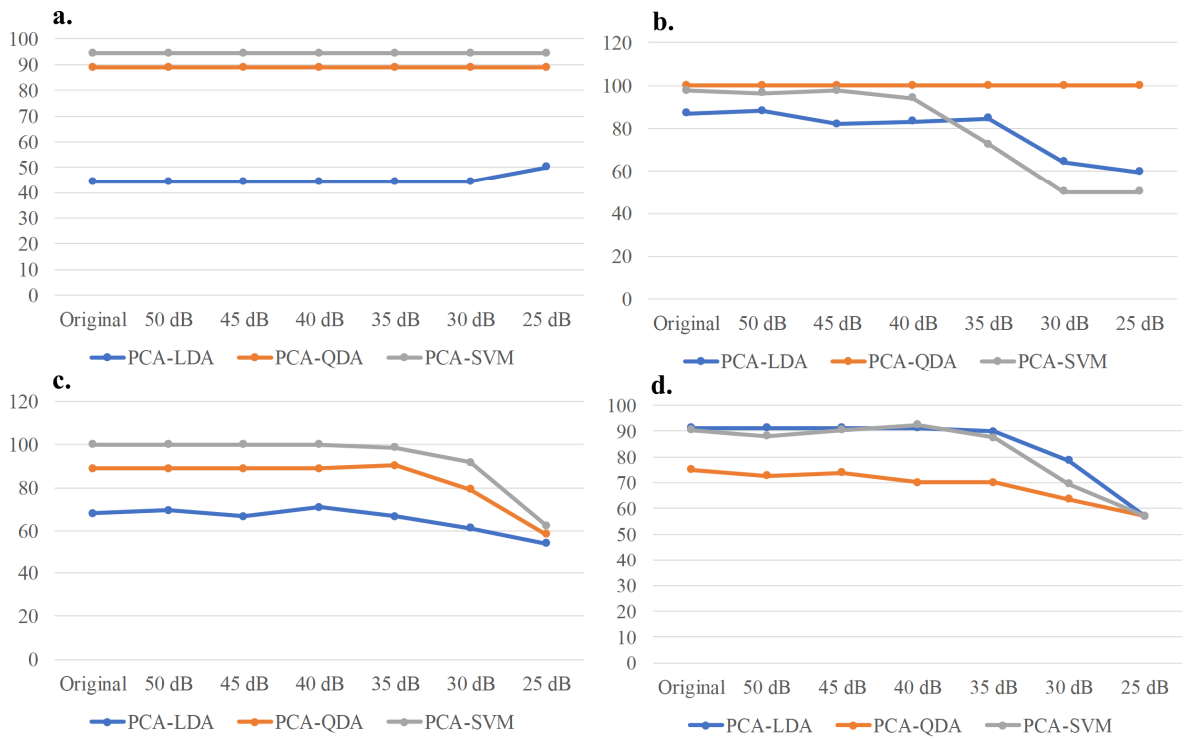
Figure 5

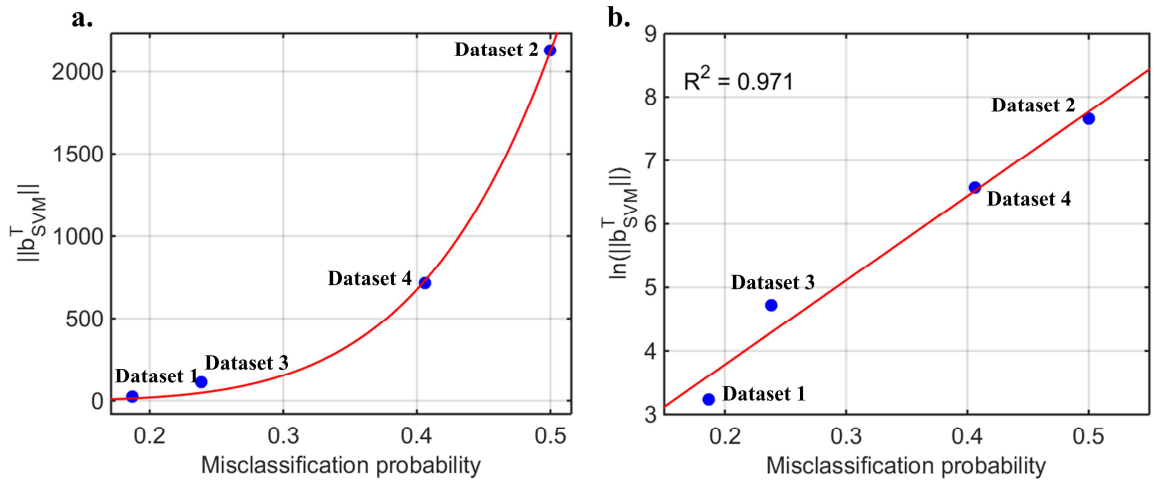





a.
Shifted Intensity
Variables
Class 2
Class 1
b.
Shifted Absorbance
Wavenumber (cm-1)
Class 2
Class 1
c.
Shifted Absorbance
Wavenumber (cm-1)
Class 2
Class 1
d.
Shifted Absorbance
Wavenumber (cm-1)
Class 3
Class 2
Class 1


a.
SVD
PCs
b.
RMSECV
PCs
c.
SVD
PCs
d.
RMSECV
PCs
e.
SVD
PCs
f.
RMSECV
PCs
g.
SVD
PCs
h.
RMSECV
PCs





Highlights:

- Misclassification probability calculation based on bootstrapping for classification
- PCA-LDA, PCA-QDA and PCA-SVM models evaluated
- Four datasets (1 simulated and 3 real applications) tested
- The misclassification probability correlates with model robustness