Received: 9 October 2019 | Revised: 12 November 2019 | Accepted: 15 November 2019

DOI: 10.1002/cem.3202

SPECIAL ISSUE - RESEARCH ARTICLE

WILEY Journal of CHEMOMETRICS

Modified PCA and PLS: Towards a better classification in Raman spectroscopy–based biological applications

Shuxia Guo[1,2] | **Petra Rösch**[1,2] | **Jürgen Popp**[1,2] | **Thomas Bocklitz**[1,2]

[1]Leibniz Institute of Photonic Technology (Leibniz-IPHT) Jena, Member of Leibniz Research Alliance 'Health Technologies', 07745 Jena, Germany

[2]Institute of Physical Chemistry and Abbe Center of Photonics, University of Jena, 07743 Jena, Germany

Correspondence
Thomas Bocklitz, Leibniz Institute of Photonic Technology (Leibniz-IPHT) Jena, 07745, Jena, Germany.
Email: thomas.bocklitz@uni-jena.de

Funding information
Bundesministerium für Bildung und Forschung, Grant/Award Number: Uro MDD (FKZ 03ZZ0444J) of 3Dsensation; Intersept; Deutsche Forschungsgemeinschaft, Grant/Award Number: BO4700/4-1; China Scholarship Council, Grant/Award Number: Scholarship from China Scholarship Council (CSC) f; BMBF, Grant/Award Numbers: FKZ 01ET1701 and FKZ 13N13852

Abstract

Raman spectra of biological samples often exhibit variations originating from changes of spectrometers, measurement conditions, and cultivation conditions. Such unwanted variations make a classification extremely challenging, especially if they are more significant compared with the differences between groups to be separated. A classifier is prone to such unwanted variations (ie, intragroup variations) and can fail to learn the patterns that can help separate different groups (ie, intergroup differences). This often leads to a poor generalization performance and a degraded transferability of the trained model. A natural solution is to separate the intragroup variations from the intergroup differences and build the classifier based on merely the latter information, for example, by a well-designed feature extraction. This forms the idea of this contribution. Herein, we modified two commonly applied feature extraction approaches, principal component analysis (PCA) and partial least squares (PLS), in order to extract merely the features representing the intergroup differences. Both of the methods were verified with two Raman spectral datasets measured from bacterial cultures and colon tissues of mice, respectively. In comparison to ordinary PCA and PLS, the modified PCA was able to improve the prediction on the testing data that bears significant difference to the training data, while the modified PLS could help avoid overfitting and lead to a more stable classification.

KEYWORDS

factor methods, feature extraction, PCA, PLSR, Raman spectroscopy

1 | INTRODUCTION

Raman spectroscopy saw dramatic growth in biological applications in the last two decades, thanks to its features: label-free, noninvasive, and almost insensitive to water.[1,2] Raman-based studies have covered a large variety of biological fields, including but not limited to toxicology and forensics,[3] microbiology,[4] drug discovery,[5,6] metabolic investigations,[7] and even in vivo detection.[8] Notably, Raman spectroscopy has found its place in process analytical technology (PAT); for example, it was employed to optimize ethanol fermentation.[9] Apart from the technical improvement in Raman spectroscopy, the booming of these applications owes to a large extent to chemometrics, where the Raman signals are translated

into high-level chemical and/or biological information of interest.[10] Such translation is usually conducted with a statistical model that takes in the Raman signals and outputs the high-level information such as biological compositions, disease levels, or bacterial types. This model is usually built on a certain number of known samples (ie, training data) and then used to predict unknown samples in a future phase to obtain the requested high-level information directly.

While a regression model is needed in quantitative applications, chemometrics in Raman-based biological applications boils down mostly to a classification task, for instance, to distinguish alterated against healthy cells/tissues or to differentiate different bacterial species/strains. To classify Raman spectra in biological applications is challenging because of multiple reasons. First of all, the high dimensionality of Raman spectra can hinder a classification, especially if the number of samples is limited, as it is often encountered in biological applications. Furthermore, spectral variations caused by biological changes of interest (ie, intergroup variations) are very subtle and hard to detect. More importantly, samples of the same biological group may exhibit different spectral features because of changes in spectrometers, environmental conditions, and cultivation conditions. The subtle intergroup spectral differences are easily overwhelmed by intragroup spectral variations, which can degrade the stability and generalizability of a classifier. With this regard, separating the intergroup variations from the unwanted changes can help improve the performance of the classification. This can be achieved with the help of feature selection, ie, to choose a subset of variables that are more different between groups and less sensitive to experimental changes.[11,12] In most cases, however, feature selection is time consuming and its performance relies largely on the procedure that is used for feature searching.

An alternative is feature extraction, which aims to transform data into a new space of lower dimension without losing key information. The new space is calculated to represent the major pattern of the dataset. However, it is not natural for feature extraction approaches to extract intergroup variations. This can be explained using two examples, the principal component analysis (PCA) and partial least squares (PLS). PCA projects datasets along the direction representing the largest sources of variance and captures the representative properties of a dataset. It is an unsupervised method, thus the resulting components do not necessarily reveal patterns that are directly related to the classification. The first principal components represent the largest variances in the dataset, which can very likely be the intragroup variations. To select the components directly related to the classification is possible but tedious. The situation is slightly better with PLS. As a supervised method, PLS seeks for projections so that the different groups show the best separation in the new data space. However, PLS is not completely robust to the intragroup variations. The influence of the unwanted variations on PLS is manifested by the fact that PLS is easily overfitted. Also, its performance largely depends on the number of components to be used in the subsequent analysis. An optimization procedure is cumbersome and remains an open issue.

To improve the quality of the extracted features and hence the performance of the classification, both PCA and PLS were modified in this contribution. The investigation was conducted on the basis of two Raman spectral datasets, which were measured from mice colon tissues and bacteria, respectively. The proposed methods were verified by the results of the classification models, and the performance was compared with their ordinary counterparts. It was proven that the modified PCA (mPCA) could help build a model with better generalizability, while the modified PLS (mPLS) is able to improve the stability of the model and avoid overfitting.

2 | EXPERIMENTAL AND METHODS

2.1 | Datasets

2.1.1 | Mice colon dataset

This dataset was measured from colon and rectum tissues of mice. The measurement was done in two cases: fully prepared samples from 47 individuals and biopsy samples from 97 individuals. Raman microspectroscopy was conducted in a grid-scan manner on each tissue sample, resulting in a certain number of Raman spectra for each sample. The number of spectra differed from sample to sample, depending on the size of the tissue. The annotation of each Raman spectrum was determined by a pathologic inspection on hematoxylin and eosin (HE)–stained tissue sections. We defined seven classes for the annotation: normal epithelium, hyperplasia, adenoma, and carcinoma, which indicate the states of the adenoma-carcinoma sequence. All other tissue types except the epithelium were annotated as "morphology." The annotation "spectroscopy" denotes substrate background or spectroscopic artifacts like burning or fluorescence, while "question" means an unambiguous annotation was impossible. The details of the sample preparation, Raman spectroscopy, and annotation can be found in Vogler et al.[13] In Table 1, we listed the number of samples belonging to the four

TABLE 1 Sample size of the mice colon dataset summarized after all preprocessing steps

		Normal	Hyperplasia	Adenoma	Carcinoma	In total
Prepared	#mice	47	12	37	14	47
	#scans	219	63	150	53	485
Biopsy	#mice	76	17	26	2	97
	#scans	171	21	43	2	237

adenoma-carcinoma states. To make it clear, we will refer to data or samples from the same mouse/individual as one replicate in the following text. For the analysis in this contribution, we used only the Raman spectra from normal, adenoma, and carcinoma groups. In addition, we combined adenoma and carcinoma as "abnormal" group and built a binary classification: normal against abnormal.

2.1.2 | Bacteria dataset

The dataset is composed of Raman spectra measured in single-cell mode from six bacterial species: *Escherichia coli* DSM 423 (*E. coli*), *Klebsiella terrigena* DSM 2687 (*K. terrigena*), *Pseudomonas stutzeri* DSM 5190 (*P. stutzeri*), *Listeria innocua* DSM 20649 (*L. innocua*), *Staphylococcus warneri* DSM 20316 (*S. warneri*), and *Staphylococcus cohnii* DSM 20261 (*S. cohnii*). All species were independently cultivated in nine replicates. The sample preparation and Raman spectroscopy has been described in one of our previous studies.[14] The sample size of each species was summarized in Table 2. The mean spectra of the six species are shown in Figure 2A.

2.2 | Data analysis

2.2.1 | Modified principal component analysis

Given a dataset $\mathbf{X} \in \mathbb{R}^{m,n}$ composed of l groups and k replicates, the loadings $\mathbf{V}$ of mPCA are calculated by a singular value decomposition (SVD) on $\sum_{\mathbf{X}}^{'}$, as is given by Equation (1). Herein, $\sum_{\mathbf{X}}^{'}$ is calculated by Equation (2), where $\sum_{\mathbf{X}}$ represents the covariance matrix of all spectra in the dataset, and $\sum_{\text{sub}}$ gives the covariance matrix from intragroup variances. By subtracting $\sum_{\text{sub}}$ from $\sum_{\mathbf{X}}$, the obtained principal components $\mathbf{V}$ are supposed to indicate merely the variances of interest, ie, the intergroup differences. The $\sum_{\text{sub}}$ is formulated as Equation (3) in our study, where the three items represent the interreplicate, intrareplicate, and intergroup variations, respectively. In particular, the covariance matrix $\sum_{br}^{g}$ is calculated from the mean spectra of each replicate belonging to the gth group. $\sum_{wr}^{i,g}$ is the covariance matrix of the ith replicate belonging to the gth group. $\sum_{bg}$ is the covariance matrix from the mean spectra of each group.

$$\sum_{X}^{'} = USV', \tag{1}$$

$$\sum_{X}^{'} = \sum_{X} - \sum_{\text{sub}}, \tag{2}$$

$$\sum_{\text{sub}} = \frac{1}{l}\sum_{g=0}^{l}\sum_{br}^{g} + \frac{1}{l \cdot k}\sum_{g=0}^{l}\sum_{i=0}^{k}\sum_{wr}^{i,g} - \sum_{bg}. \tag{3}$$

TABLE 2 Sample size of the bacteria dataset. The information was summarized based on preprocessed spectra, excluding the outliers

	Gram negative (Gram−)			Gram positive (Gram+)		
Species	***Escherichia coli***	***Pseudomonas stutzeri***	***Klebsiella terrigena***	***Listeria innocua***	***Staphylococcus cohnii***	***Staphylococcus warneri***
#Spectra	447	458	451	455	449	448

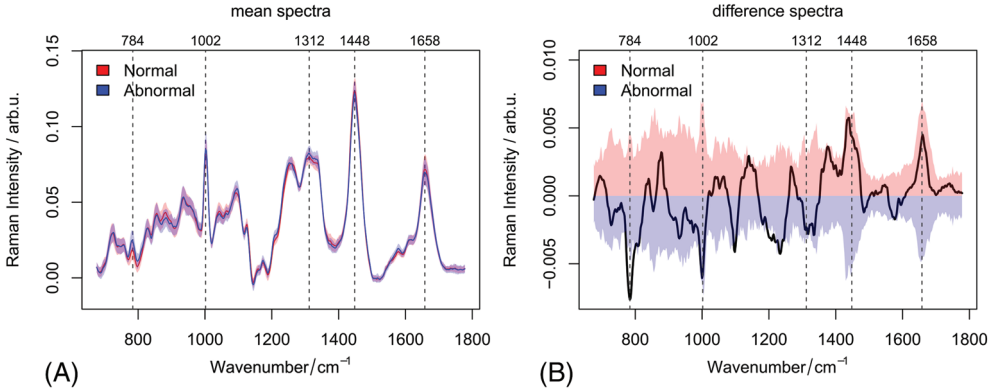

FIGURE 1 Mean spectra, variances, and difference spectrum of the mice colon data. The mean and difference spectra are shown as solid lines in both plots, while the variances are given as the red/blue shade. The key Raman bands were marked by dash lines. The mean spectra of the two groups are very similar to each other. Their differences, as is shown in (B), are mostly dominated by the intragroup variances. The only dominating intergroup difference was observed for the Raman band 784 cm^{-1}

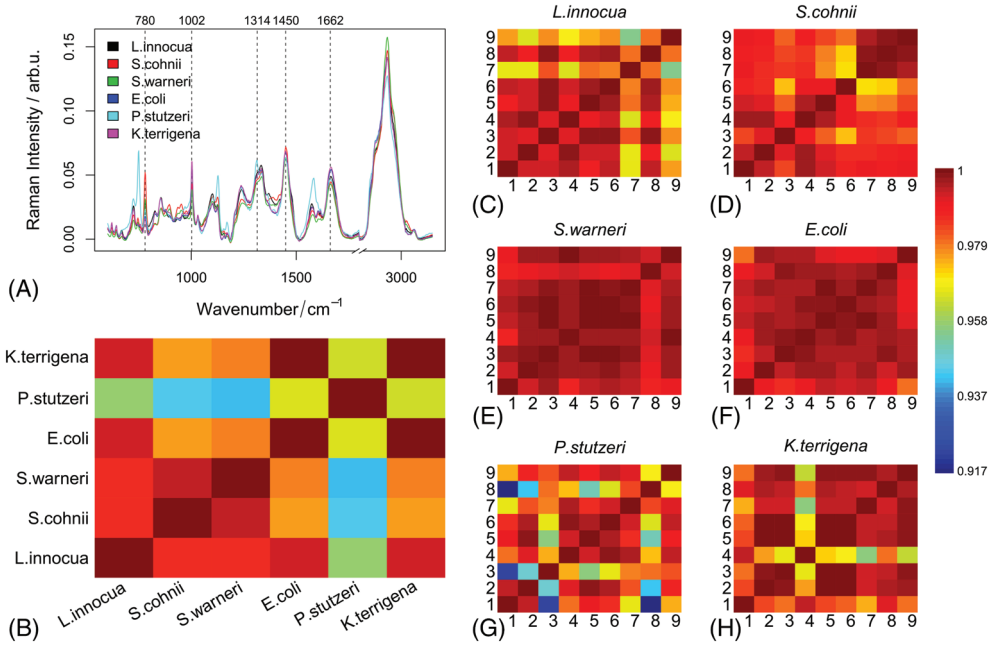

FIGURE 2 Mean spectra and correlation coefficients of bacteria dataset. A, Mean spectra of each species calculated from all batches. The key Raman bands were marked by dash lines. B, Intergroup correlation coefficients calculated from the mean spectrum of each species. C-H, Interreplicate correlation coefficients of different species, calculated from the mean spectrum of each batch from the same species. In particular, species *Listeria innocua*, *Staphylococcus cohnii*, and *Staphylococcus warneri* are Gram positive (Gram+) while the other three species are Gram negative (Gram−). The intragroup variations surpass the intergroup differences if the interreplicate coefficients are higher than the intergroup coefficients and vice versa. Accordingly, the differences between Gram+ and Gram− are marginally higher than the intrareplicate variations. The two variations are comparable for the three Gram+ species. The intragroup variations are larger than the intergroup variations for the Gram− species, especially in the case of *Escherichia coli* and *Klebsiella terrigena.*

2.2.2 | Modified partial least squares

Similar to mPCA, the idea of mPLS is to calculate projection vectors that represent merely the variations of interest, ie, the intergroup variations in this study. As the first step, the matrix $\sum_{\text{sub}}$ was decomposed by a SVD to obtain the loadings ($\mathbf{L}_{\text{sub}}$) representing the intragroup variations. The calculation of mPLS was based on the SIMPLS algorithm,[15] by orthogonalizing the projection vectors ($\mathbf{P}$) against $\mathbf{L}_{\text{sub}}$ during the iteration. The orthogonalization of $\mathbf{P}(j)$ against $\mathbf{L}_{\text{sub}}(k)$, according to the Gram-Schmidt method, is shown in Equation (4). The j,k represent the indices of column vectors in $\mathbf{P}$ and $\mathbf{L}_{\text{sub}}$, respectively. In the end, we could obtain the score vectors of mPLS based on the projection vectors $\mathbf{P}_{ot}$.

$$\mathbf{P}_{ot}(j) = \mathbf{P}(j) - \left(\frac{\mathbf{P}(j)\cdot\mathbf{L}_{\text{sub}}(k)}{\mathbf{L}_{\text{sub}}(k)\cdot\mathbf{L}_{\text{sub}}(k)}\right)\mathbf{L}_{\text{sub}}(k);\ j,\ k = 1,\ 2,\ \cdots. \tag{4}$$

2.2.3 | Method validation

The performance of mPCA and mPLS were compared with that of the ordinary PCA and PLS on the basis of the two above-mentioned datasets. All data analyses were conducted with in-house written scripts based on R language.[16] The analysis of the mice colon data started from the spikes removal for each single spectrum using a derivative-based algorithm, followed by a wavenumber calibration on the basis of a standard material 4-acetamedophenol.[17] The fluorescence baseline was then corrected by the alternative least squares (ALS).[18] All spectra were truncated to 675 to 1800 cm^{-1} after baseline correction, and a vector normalization (ie, l_2 norm) was conducted. We did not perform outlier detection but only excluded the spectra belonging to groups "morphology," "spectroscopy," and "question" according to the annotation process done in Vogler et al.[13] After all these preprocessing steps, we averaged the spectra belonging to the same group for each scan. This averaging procedure may result in one, two, three, or four spectra out of one scan, depending on how many groups were there in this scan. The sample size given in Table 1 was summarized on the basis of the average spectra. The mean and difference spectra after preprocessing are plotted along with the variances in Figure 1 for normal and abnormal groups. The two groups were differentiated with a binary classifier composed of (m)PCA/(m)PLS and linear discriminant analysis (LDA). Only the prepared samples were used for the model training; the biopsy samples were used merely as testing data. The model building was combined with a 10-fold cross-validation, in which both feature extraction and LDA were involved within the cross-validation loop (see "inside CV" in Guo et al[19]). Data from the same mouse were used exclusively in one fold. The biopsy samples were predicted every time the model was built during the cross-validation. The performance of the classification was benchmarked by the mean sensitivity, which is defined in Equation (5) for an l-group classification task.

$$\overline{s} = \frac{1}{l}\sum_{i=1}^{l} s(i). \tag{5}$$

In the case of the bacteria data, the spikes were detected by comparing two repeated measurements, and their values were replaced with the lower intensities of the two spectra. The two spectra without spikes were averaged so that we would get a single spectrum for each cell. Subsequent wavenumber calibration and baseline correction were conducted with the same approaches as for the mice colon data. The resulting spectra were vector-normalized after removing the silent region (1750 to 2800 cm^{-1}). In the end, we excluded three Raman spectra from further analysis because of obvious burning artifacts (Figure S1); no additional outlier detection was performed. The sample size was summarized in Table 2 after all preprocessing steps. The mean spectra of each species were plotted in Figure 2A. The dataset was then fed into a two-layer classification model: a binary classifier in the first layer to separate Gram positive (Gram+) against Gram negative (Gram−), following two three-group classifiers in the second layer to differentiate the three Gram+ and Gram− species, respectively. We used (m)PCA/(m)PLS in combination with an LDA for all the three classifiers. The classification was performed in combination with a leave-one-replicate out cross-validation in "inside CV" mode.[19] The performance was benchmarked by the mean sensitivity (Equation 5).

3 | RESULTS AND DISCUSSION

3.1 | Mice colon data

One of the requirements for the proposed method to work is that the intergroup variations are not significantly sacrificed by removing the intragroup variations. To verify this point, we checked the loading vectors of the matrix $\sum_{\text{sub}}$ shown in Figure 3 and compared them with the variance and difference spectrum given in Figure 1B. The Raman bands at 1002, 1312, 1448, and 1658 cm^{-1}, where we see dominating intragroup variances (shades in Figure 1B), are visible in the two loadings. This demonstrated that the intragroup variations are captured well by $\sum_{\text{sub}}$. In addition, there is no significant signal at 784 cm^{-1} in either loading, where the intergroup differences were larger than the intragroup

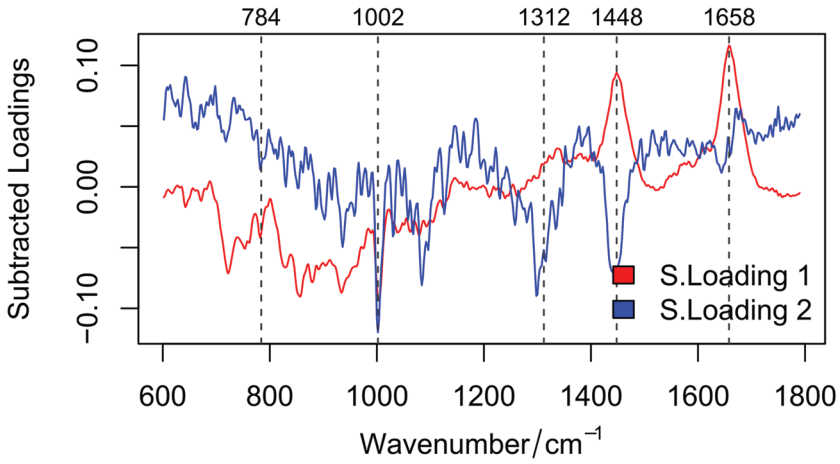

FIGURE 3 The first two loadings that are calculated from the matrix $\sum_{sub}$. The peaks at 1002, 1312, 1448, and 1658 cm^{-1} represent high intragroup variances, which is consistent with the variances shown in Figure 1. The peak at 784 cm^{-1}, which features major intergroup differences, is not visible in either loading

variances. Although it is too early to say whether the intergroup variations are completely retained, we can already conclude that the major features related to the intergroup differences are not sacrificed.

The influence of the modification on the calculated loadings is indicated by the first two components of the ordinary PCA and the mPCA plotted in Figure 4. Comparing with the ordinary PCA, the band located around 784 cm^{-1} was significantly enhanced in the first loading of the mPCA. This is consistent with the biological knowledge that the amount of RNA/DNA, represented by the Raman band at around 784 cm^{-1}, is significantly different between normal and cancerous groups. The difference spectrum in Figure 1B also indicates this fact: the intergroup difference visibly surpasses the intragroup variances within the region around 784 cm^{-1}. In the meantime, the sharp band at around 1002 cm^{-1}, which is dominating in both loadings of ordinary PCA because of the large intragroup variances, is suppressed in both

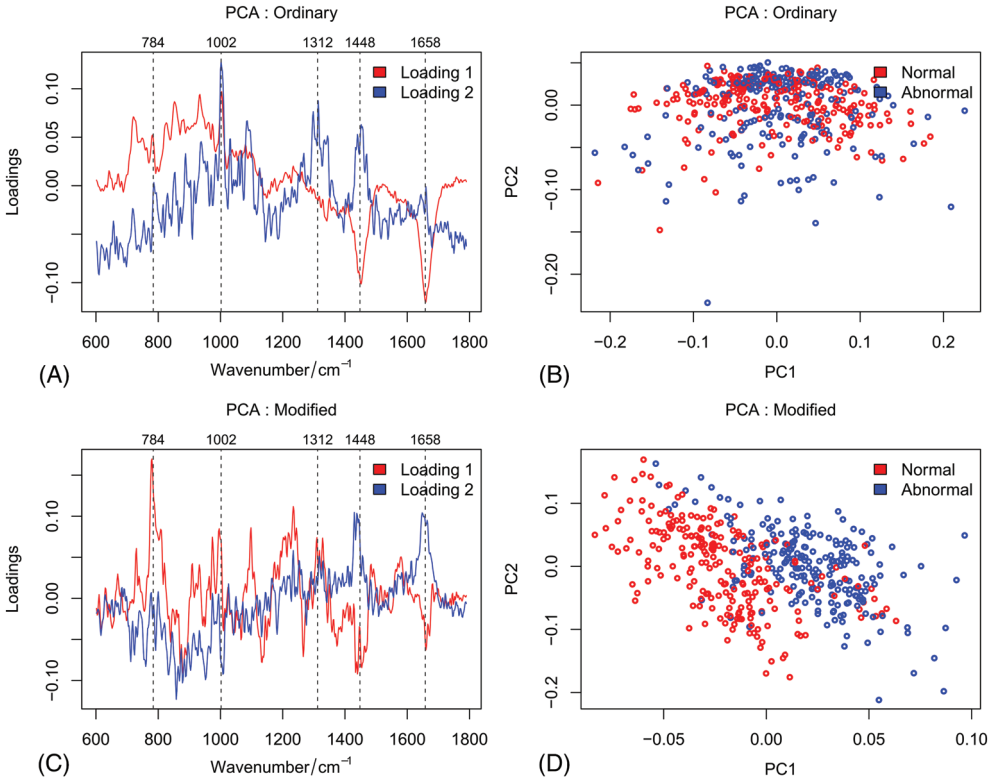

FIGURE 4 The loadings and scores of the first two components for ordinary (A,B) and modified (C,D) principal component analysis (PCA). The major intergroup difference was enhanced in modified PCA (mPCA) loadings comparing with the loadings of the ordinary PCA. The Raman bands at 1002, 1312, 1448, and 1658 cm^{-1} showing large intragroup variations are suppressed but still present in the mPCA loadings. In comparison of the scores plots in (B) and (D), a clearer separation between the normal and abnormal groups is obtained with mPCA.

loadings of the mPCA. Nonetheless, the features at 1002, 1312, 1448, and 1658 cm^{-1} are still present in the mPCA loadings. This indicates most probably that the amount of protein differs as well between normal and cancer cells. On the basis of these facts, it is reasonable to say that the mPCA suppresses features bearing large intragroup variations but not significantly sacrifices intergroup differences. Consequently, we could see a better separation of the two groups with the mPCA than the ordinary PCA in the score plots of the first two components.

We could verify the influence of the modification on the PLS in a similar way. As it is shown in Figure S2(A,B), the Raman band at around 784 cm^{-1} was already well represented in the projection vectors and the two groups were well separated by the ordinary PLS. This is reasonable considering the fact that PLS is a supervised method and extracts features mostly related to the intergroup differences. Nonetheless, we could still see the difference between the ordinary PLS and mPLS. The bands at around 1002, 1312, 1448, and 1658 cm^{-1} were still present but suppressed in the case of mPLS (Figure S2C), indicating a reduced contribution of intragroup variations in the mPLS model. In comparison of Figure S2B,D, we could see more compact cluster of the same group for mPLS, even though the separability between the groups was not improved.

Besides the output of the mPCA and mPLS, we tested further their performance with the binary classification of the mice colon data. In particular, the prepared samples were predicted by a 10-fold cross-validation while the biopsy samples were predicted 10 times during the cross-validation using the model built on nine out of the 10 folds. The mean sensitivity for the biopsy samples was calculated for each of the 10 predictions and averaged. The biopsy samples were only predicted, because of two reasons. First, this approach aligns with the diagnostic workflow in practice. Second, a well-known fact in real diagnostics is that the prepared and biopsy samples are significantly different and a model transfer is very likely needed to predict biopsy samples with a model built on prepared samples. With this experiment, we aim to check if the prediction on biopsy samples can be improved by modifying the PCA and PLS.

The mean sensitivity of the predictions is shown in Figure 5, in which the results from ordinary and modified PCA/PLS were plotted in blue and red color, respectively. The prediction of the prepared samples was comparable for ordinary PCA and mPCA. A possible reason is that the classifier learns well about the intergroup differences based on the sufficiently large number of mice. A modification of PCA does not further improve the performance. Nonetheless, the classifier became more tolerant to the differences between the prepared and biopsy samples, and the prediction

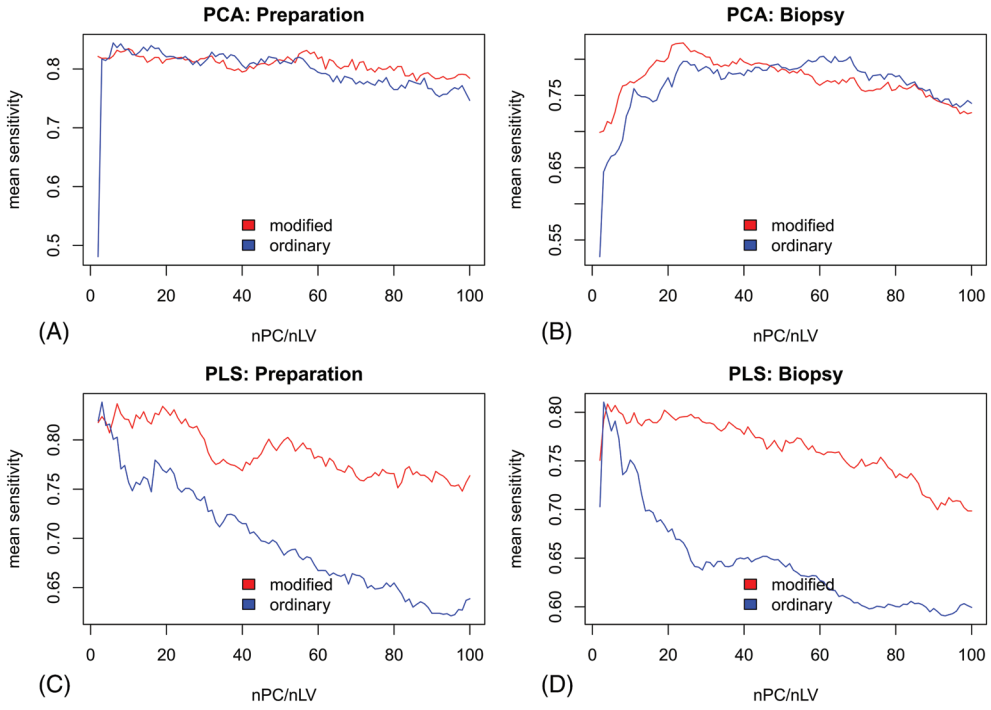

FIGURE 5 Mean sensitivity of mice colon dataset as a function of number of components (nPC/nLV), using (m)PCA/(m)PLS as the feature extraction methods. The "Preparation" and "Biopsy" denote the results of cross-validation and prediction of the biopsy samples, respectively. The prediction on preparation data were comparable for the modified and ordinary PCA, while an improvement was seen for mPCA to predict the biopsy samples. The results of PLS was marginally higher for mPLS than ordinary PLS, especially with higher number of components. mPCA, modified PCA; mPLS, modified PLS; PCA, principal component analysis; PLS, partial least squares.

of the biopsy samples was improved to almost the same level as for the prediction on the prepared samples. This demonstrates a better generalization/transferability of the model from the prepared samples to biopsy samples. The results of PLS tells a different story. The ordinary PLS and the mPLS are comparable in terms of their highest mean sensitivity over different number of components (nLV). This can be explained by the fact that the mPLS does not improve the separability of the two groups, as it was discussed previously. However, the advantage of the mPLS is demonstrated when it comes to higher values of nLV. The mean sensitivity drops sharply with increased nLV after the peak point in the case of the ordinary PLS. The decrease was much slower for the mPLS. That is to say, the mPLS-based classification is more stable and less prone to overfitting because of the lower influence from intragroup variations.

3.2 | Bacteria data

The performance of the mPCA and mPLS was further verified on the basis of the bacteria data. The classification was achieved using a two-layer model, as it was described in the previous section. With almost equal number of spectra for each species and replicate, we could discuss the results taking the different levels of intragroup and intergroup variations into account. To do so, we visualized the mean sensitivity of not only the two-layer classification (Figure 6A,B) but also the two 3-group classifiers in the second layer (Figure 6C-F). In addition, we calculated the Pearson correlation

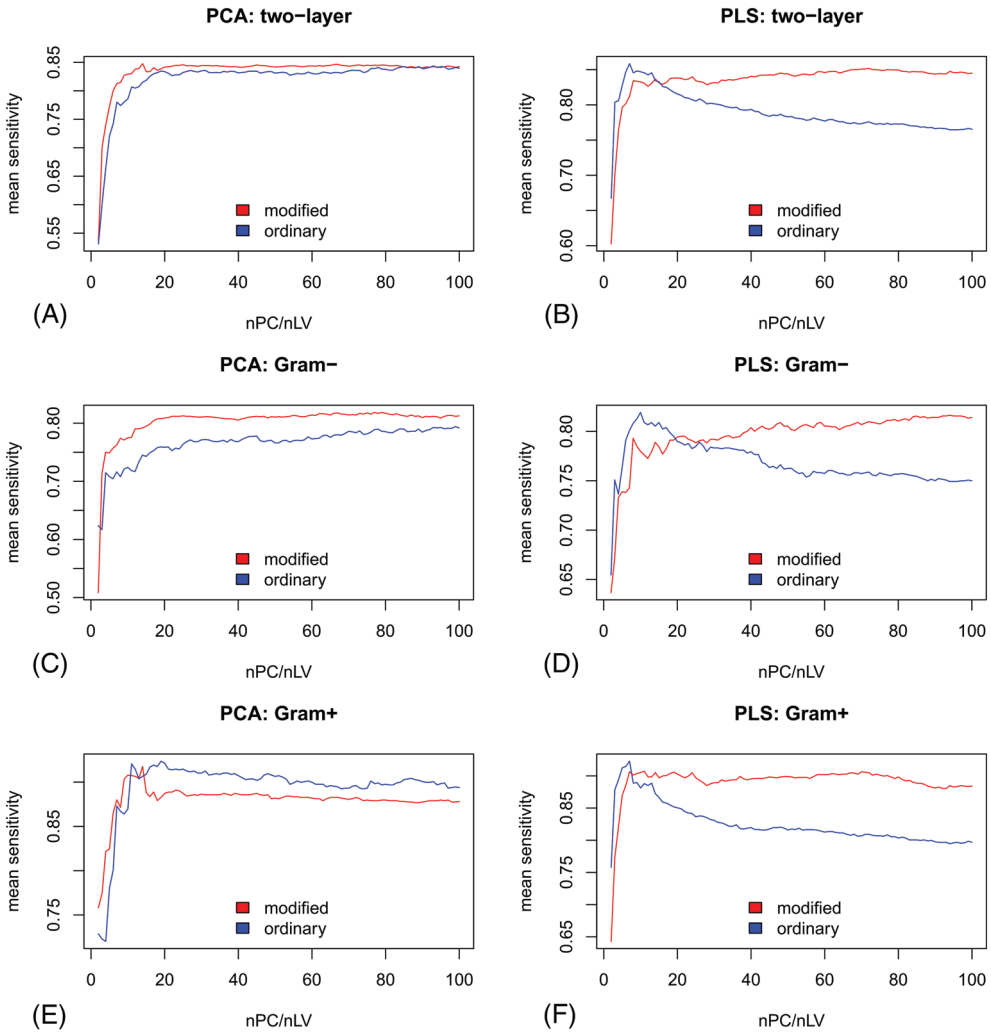

FIGURE 6 Mean sensitivity of the hierarchical models for the bacteria dataset as a function of the number of components (nPC/nLV). The results of ordinary and modified PCA/PLS are plotted in blue and red color, respectively. The prediction was better for mPCA than ordinary PCA in the case of two-layer model and the classification of Gram− species, while ordinary PCA is better at differentiating the Gram+ species. The mean sensitivity is not improved by mPLS compared with the ordinary PLS, but it decreases much slower with nLV in the case of mPLS. mPCA, modified PCA; mPLS, modified PLS; PCA, principal component analysis; PLS, partial least squares.

coefficients as a metric of the intergroup and intragroup variations. The intergroup correlation (Figure 2B) was computed from the mean spectra of different species, while the intragroup correlation (Figure 2C-H) was calculated for each species using the mean spectra of different replicates.

Apparently, the performance of mPCA varies among different classifications, as can be seen from the left column in Figure 6. A better prediction was achieved with mPCA for the two-layer model (Figure 6A) and the separation of the Gram− species (Figure 6C). However, the mPCA was inferior to the ordinary PCA in terms of the classification of the Gram+ species (Figure 6E). To explain this issue, we calculated the correlation coefficients between different species and between replicates, as were shown in Figure 2B-H. It is obvious from the plots that the intergroup variations are lower than the intragroup variations for the Gram+ species. The separation of the Gram+ species was almost tolerant to the intragroup variations. A further modification of PCA is hence unnecessary and can negatively impact their performance. The situation is different for the Gram− species, where the intergroup variations are too small to compete the intragroup variations. It is helpful to remove the intragroup variations from the calculation. This fact is more obvious when it comes to the binary classification between *E. coli* and *K. terrigena* (E-K). In this case, the intergroup variations are highly dominated by the intragroup variations (see Figure 2B,F,H), and the mean sensitivities (Figure S3) were marginally better for the mPCA than for the ordinary PCA. On the basis of these results, we would like to stress that the mPCA should be used with caution. It is recommended that the intergroup and intragroup variations be compared at the first place. The mPCA is more suited if the intergroup differences are dominated by the intragroup variations.

In comparison with PCA, the competition between intergroup and intragroup variations is less influential for the mPLS. The performance was seen similar in the three classification tasks (Figure 6B,D,F). This originates from the fact that PLS is a supervised approach and extracts features that are mostly related to the separation. The removal of the intragroup variations does not significantly influence the separability but only improves the stability of the model, as it was discussed previously. The results in Figures 6 and S3 once again proved this fact, as the highest mean sensitivity was comparable for mPLS and ordinary PLS. Nonetheless, the mPLS becomes more stable and less overfitted in general. This is especially obvious for the E-K binary classification (Figure S3B), in which the ordinary PLS degrades sharply with higher nLV because of the significant intragroup variations. The prediction became more stable over nLV by removing the intragroup variations in the case of mPLS.

Besides the competition between intragroup and intergroup variations, another issue may also influence the performance of mPCA and mPLS: the sample size. Here, we refer it to especially the number of replicates included in the training data. It is required that the replicates form a good representative of intragroup variations. Otherwise, the intragroup variations cannot be well extracted or removed properly from the calculation. The influence of sample size in statistical modeling falls into the field of sample size planning and was investigated in one of our latest studies.[14]

4 | CONCLUSION

In this study, we modified two commonly used feature extraction methods, PCA and PLS, in order to deal with the large intragroup variations caused by experimental changes from replicate to replicate. With this modification, we could improve the generalization performance or stability of the classification models. In particular, we modeled the intragroup variations into a matrix ($\sum_{sub}$) composed of variations between- and within-replicates excluding the intergroup differences. The PCA and PLS was conducted so that the extracted loading/projection vectors are orthogonal to the intragroup variations, which results in the features represent merely the variations of interest. The proposed methods were tested with two Raman spectral datasets measured from mice colon tissues and bacteria, respectively. Their performance was benchmarked by the mean sensitivity of different classification tasks in combination with a leave-one-replicate cross-validation. With the mice colon data, we could observe a better transferability/generalization of the classifier from the prepared to the biopsy samples. It was found using the bacteria dataset that the performance of mPCA is dependent on the ratio between intragroup and intergroup variations. The mPCA is more suited if the intragroup variations are larger than the intergroup differences. Otherwise, the modification may degrade the classification. This was different for PLS. Despite no significant improvement for the separation between groups, mPLS was found to yield more compact group clusters in the scores space. This helps improve the stability of the classification and avoid overfitting of the model, as was demonstrated in this study. In particular, we conducted the classification in all cases with nPC/nLV ranging up to 100 for a more conclusive comparison between ordinary and modified PCA/PLS. However, in practice, classification and regression models should be constructed with lower nPC/nLV to increase their robustness.

ACKNOWLEDGEMENTS

The authors acknowledge the funding of the project Uro-MDD (FKZ 03ZZ0444J) from Bundesministerium für Bildung und Forschung (BMBF) via 3Dsensation and the project BO4700/4-1 from Deutsche Forschungsgemeinschaft (DFG) as well as the projects Intersept (FKZ 13N13852) and CarbaTech (FKZ 01ET1701) from the Bundesministerium für Bildung und Forschung (BMBF). The scholarship from China Scholarship Council (CSC) for S.G. is highly appreciated.

ORCID


Thomas Bocklitz https://orcid.org/0000-0003-2778-6624

REFERENCES

1. Bocklitz TW, Guo S, Ryabchykov O, Vogler N, Popp J. Raman based molecular imaging and analytics: a magic bullet for biomedical applications!? *Anal Chem*. 2016;88(1):133-151.
2. Kong K, Kendall C, Stone N, Notingher I. Raman spectroscopy for medical diagnostics—from in-vitro biofluid assays to in-vivo cancer detection. *Adv Drug Deliv Rev*. 2015;89:121-134.
3. de Oliveira Penido CAF, Pacheco MTT, Lednev IK, Silveira L Jr. Raman spectroscopy in forensic analysis: identification of cocaine and other illegal drugs of abuse. *Journal of Raman Spectroscopy*. 2016;47(1):28-38.
4. Lorenz B, Wichmann C, Stöckel S, Rösch P, Popp J. Cultivation-free Raman spectroscopic investigations of bacteria. *Trends Microbiol*. 2017;25(5):413-424.
5. Paudel A, Raijada D, Rantanen J. Raman spectroscopy in pharmaceutical product design. *Adv Drug Deliv Rev*. 2015;89:3-20.
6. Tipping WJ, Lee M, Serrels A, Brunton VG, Hulme AN. Imaging drug uptake by bioorthogonal stimulated Raman scattering microscopy. *Chem Sci*. 2017;8(8):5606-5615.
7. Shiota M, Naya M, Yamamoto T, et al. Gold-nanofève surface-enhanced Raman spectroscopy visualizes hypotaurine as a robust anti-oxidant consumed in cancer survival. *Nat Commun*. 2018;9(1):1561.
8. Desroches J, Jermyn M, Pinto M, et al. A new method using Raman spectroscopy for in vivo targeted brain cancer tissue biopsy. *Sci Rep*. 2018;8(1):1792.
9. Hirsch E, Pataki H, Domján J, et al. Inline noninvasive Raman monitoring and feedback control of glucose concentration during ethanol fermentation. *Biotechnol Prog*. 2019;35(5):e2848.
10. Oleg R, Shuxia G, Thomas B. Analyzing Raman spectroscopic data. *Physical Sciences Reviews*. 2019;4(2).
11. Guyon I, Elisseeff A. An introduction to variable and feature selection. *Journal of machine learning research*. 2003;3(Mar):1157-1182.
12. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc B Methodol*. 1996;58(1):267-288.
13. Vogler N, Bocklitz T, Subhi Salah F, et al. Systematic evaluation of the biological variance within the Raman based colorectal tissue diagnostics. *J Biophotonics*. 2016;9(5):533-541.
14. Ali N, Girnus S, Rösch P, Popp J, Bocklitz T. Sample-size planning for multivariate data: a Raman-spectroscopy-based example. *Anal Chem*. 2018;90(21):12485-12492.
15. De Jong S. SIMPLS: An alternative approach squares regression to partial least. *Chemom Intel Lab Syst*. 1993;18:2-263.
16. Team RC. *R Foundation for Statistical Computing*. Vienna: Austria; 2013:3.
17. Dörfer T, Bocklitz T, Tarcea N, Schmitt M, Popp J. Checking and improving calibration of Raman spectra using chemometric approaches. *Zeitschrift für Physikalische Chemie*. 2011;225(6-7):753-764.
18. Liland, K.H. and Mevik, B.-H., *Baseline: Baseline Correction of Spectra*. R package version 2015.
19. Guo S, Bocklitz T, Neugebauer U, Popp J. Common mistakes in cross-validating classification models. *Anal Methods*. 2017;9(30):4410-4417.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Guo S, Rösch P, Popp J, Bocklitz T. Modified PCA and PLS: Towards a better classification in Raman spectroscopy–based biological applications. *Journal of Chemometrics*. 2020;34:e3202. https://doi.org/10.1002/cem.3202