Recognition of surface-enhanced Raman spectra of organic media based on deep learning

Lyudmila A. Bratchenko[1], Sahar Z. Al-Sammarraie[1], Elena N. Tupicova[1],*
Yulia A. Khristoforova[1], Ivan A. Bratchenko[1]

1 Samara University, Samara, Russia

[*]e-mail:shamina94@inbox.ru

The study was supported by a grant from the Russian Science Foundation No. 21-75-10097, https://rscf.ru/project/21-75-10097/

1. Introduction

Despite numerous highly efficient optical methods introduced into analysis of tissues and fluids, only a small number of them actually reach clinics. One possible reason for this phenomenon is the complexity of spectral data and the problems associated with statistical analysis of spectral datasets. This study provides a detailed demonstration of how database of complex organic media SERS spectra can be used for classification and recognition by means of multivariate analysis. Human blood serum was chosen as a complex organic medium as an object of study of practical interest.

2. Materials and methods

- Silver structures based on dried silver colloid are utilized to achieve surface enhancement of Raman scattering in the near infrared range. A silver colloid was obtained by reduction from an aqueous solution of silver nitrate with sodium citrate at the temperature of 95 °C for 20 minutes.

- For SERS analysis, each serum sample was dropped in a volume of 2 μl on aluminum foil with a layer of silver structures and dried for 30 minutes. The analysis of the serum spectral characteristics was carried out using an experimental stand consisting of a spectrometric system (EnSpectr R785, Spektr-M, Chernogolovka, Russia) and a microscope (ADF U300, ADF, China). The spectra were excited in the near infrared range using a laser module with the center wavelength of 785 nm. A 50x LMPlan Objective was used to focus the radiation on the sample and collect the scattered radiation.

- The preprocessing of raw SERS serum spectra consisted of two sequential stages: noise smoothing and normalization. Smoothing the raw spectra was performed by the Savitzky-Golay filter the with filter window width of 15, the first-order of the polynomial used for smoothing and the zero-order of derivative to take (no derivative). After noise smoothing, the spectral characteristics of the serum were normalized by means of a standard deviation of the normal variate method (SNV).

- The data were analyzed through supervised learning. The classification model was trained on a full sample (50 spectra of serum from healthy subjects vs 50 spectra of serum from the patients with cardiovascular diseases) and on a reduced sample (10 spectra of serum from the healthy subjects vs 10 spectra of serum from the patients with cardiovascular diseases). The analysis of the model stability was implemented using the k-fold cross-validation (k = 7). When constructing the models, the importance of predictors in accomplishing the classification task was assessed by means of the distribution of the variable importance (VIP) in the constructed model.

Conclusion

According to the provided example of SERS database classification , we can formulate several steps required to avoid overestimation during model construction :

1. Divide the data into the calibration and the test sets, or perform CV with a limited available data; make this division sample-sensitive;
2. Run multiple divisions of data and check the model performance during different runs;
3. Choose several metrics (not a single parameter) to evaluate model performance. Analyze RMSE even for binary classification;
4. Evaluate LV shape to discard noise contribution;
5. Select only those LVs that provide stable results and contain useful spectral information;
6. Present performance as "Mean±SD" for multiple runs or add confidence interval estimation for the presented performance;

Describe all the details of the performed analysis in the Materials Section and present all the details of the performed statistical analysis (possibly as supplementary) in order to give the readers the possibility to replicate the research findings.

3. Results of multivariate analysis

The process of spectral data classification requires separation of the dataset under analysis into the calibration (train) and the test sets. This step is crucial for verifying the stability of the resulting classification model. An important conclusion about Fig.An important conclusion about Fig. 1 is that a single split of data can provide unstable results. A classification model built on only one data split may easily become overestimated and provide overoptimistic results. 1 is that a single split of data can provide unstable results. A classification model built on only one data split may easily become overestimated and provide overoptimistic results.

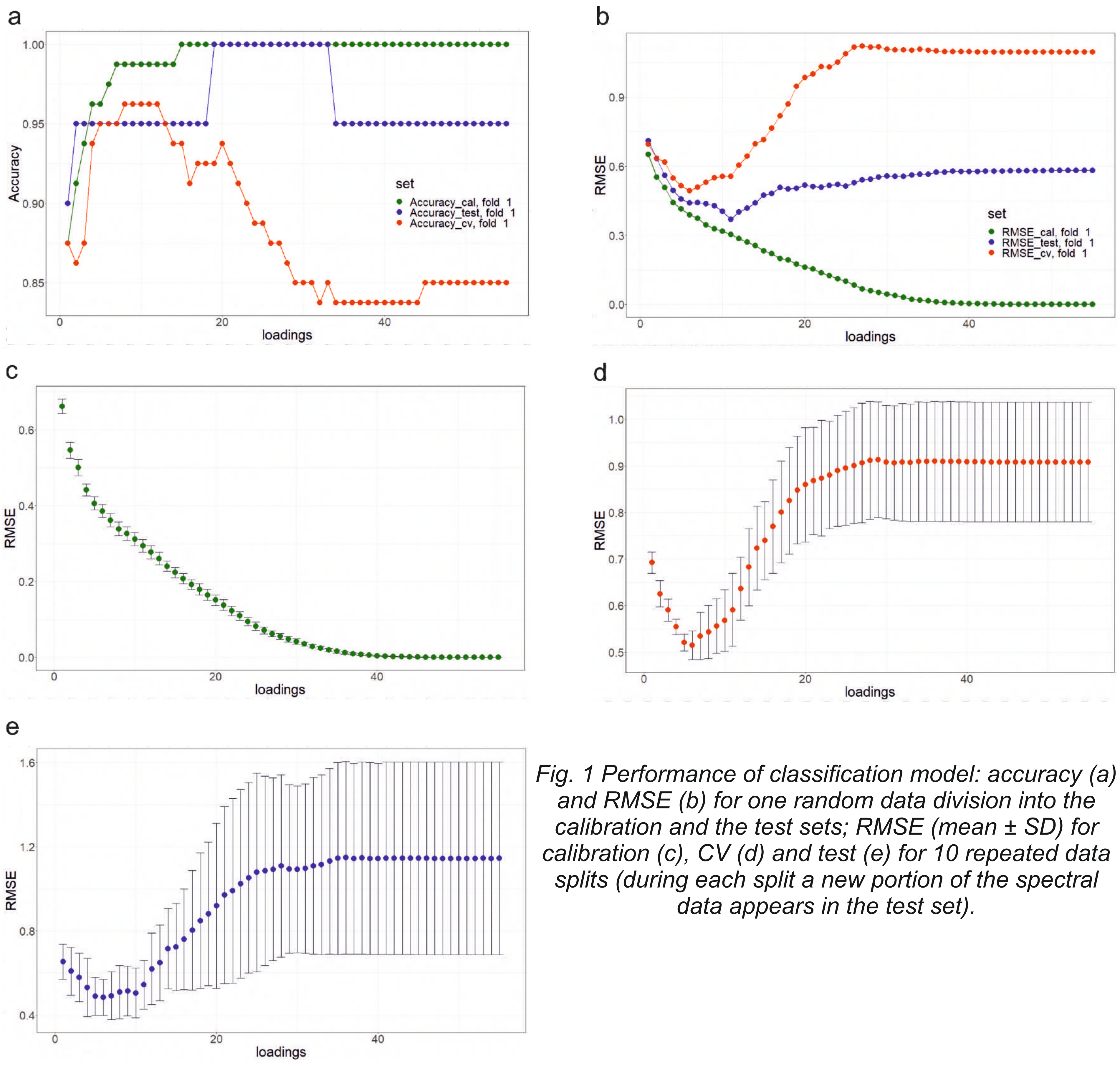

Fig. 1 Performance of classification model: accuracy (a) and RMSE (b) for one random data division into the calibration and the test sets; RMSE (mean ± SD) for calibration (c), CV (d) and test (e) for 10 repeated data splits (during each split a new portion of the spectral data appears in the test set).

Fig. 2 demonstrates LVs 1, 6, 11, 21 and 50 for one random split of the analyzed dataset . One can see that with the increased LV number, the amount of useful spectral data carried by LV decreases. An approach that highlights LVs with useful spectral data is a variable importance in projection (VIP). VIP is a weighted sum of LVs, and if the LVs' intensity is relatively small, it should not change the shape of the VIP distribution.

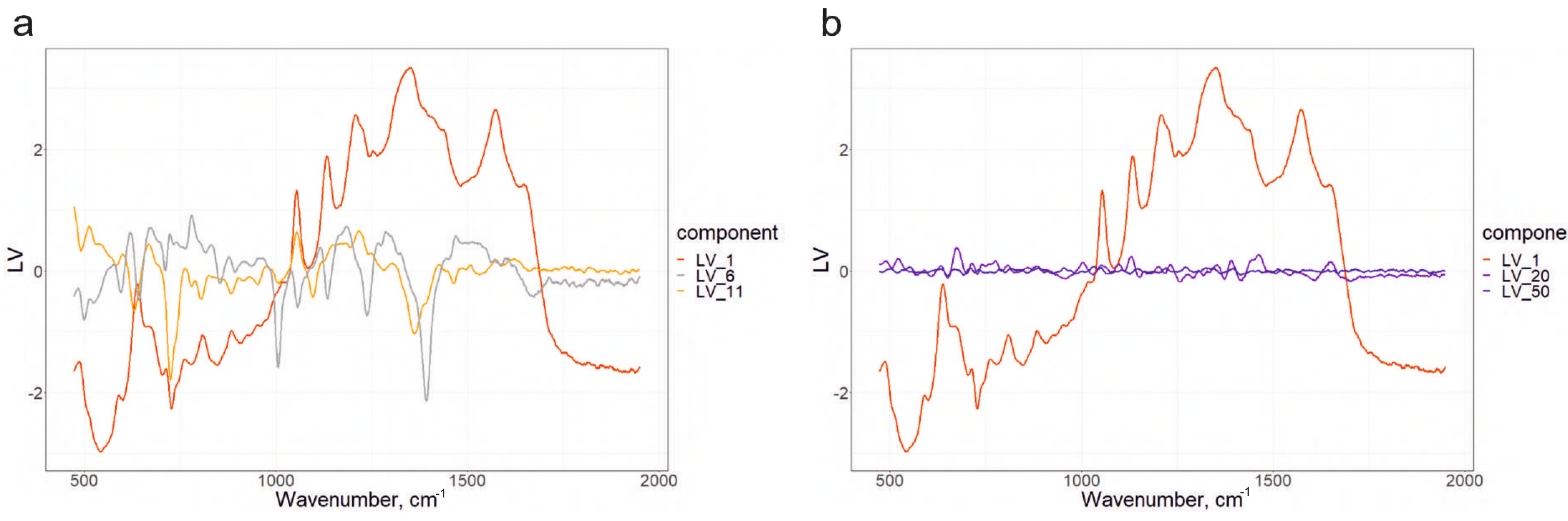

Fig. 2 LVs (loading vectors) shape for one split in PLS-DA: LV1, LV6 and LV11 (a) carry useful spectral information, while high-order LVs (e.g. LV20 and LV50) (b) carry only noises

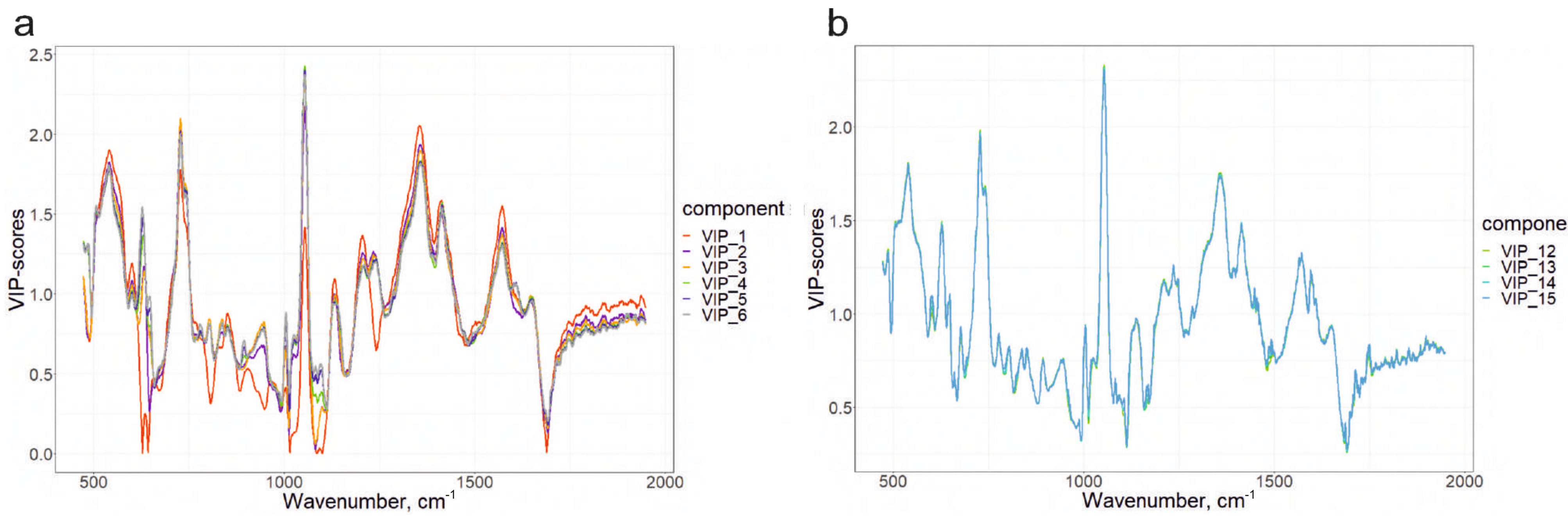

Fig. 3 VIP (variable importance in projection) for PLS-DA classification: utilizing the first LVs changes the shape of VIP distribution (a), while utilizing the high-order LVs does not change the shape of VIP distribution (b).