Reducing Temperature Calibration Error in Multivariate Analysis of Fluorescence Spectra

Mikhail Khodasevich[1*], Vladimir Aseev[2], Victor Klinkov[3], Evgenia Tsimerman[3], Darya Borisevich[1]

[1] B.I.Stepanov Institute of Physics, National Academy of Sciences of Belarus, Minsk, Belarus
[2] National Research University of Information Technologies, Mechanics, and Optics, St. Petersburg, Russia
[3] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russia
Email: m.khodasevich@ifanbel.bas-net.by

Abstract. A method has been demonstrated to reduce temperature calibration error by integrated using principal component analysis, hierarchical cluster analysis and searching combination moving window interval projection to latent structures for fluorescent spectra of Er-doped $98MgCaSrBaYAl_2F_{14}$-$2Ba(PO_3)_2$ and Yb-doped CaF_2. The consecutive and consistent use of these multivariate methods for outliers detection, forming training and test datasets and variable selection is shown to allow more than twofold reducing the root-mean-square error of temperature calibration in comparison with the application of projection to latent structures without variable selection.

Keywords: Projection to latent structures, principal component analysis, cluster analysis, calibration, fluorescence spectrum.

1 Introduction

Currently there are several types of fluorescent sensors for temperature calibration. They can be based on the active medium fluorescence band shift [1], fluorescence decay time change [2] and fluorescence intensity change [3] or fluorescence intensities ratio from two thermally-coupled energy levels [4] or non-coupled levels of doping ions [5], respectively. From the fluorescence spectra, only measurements at a small number of wavelengths are usually used to calibrate the temperature. In [6], we showed the possibility of temperature calibration while using the multivariate method of spectra analysis – projection to latent structures (PLS) [7]. Considering the upconversion fluorescence of erbium-doped lead-fluoride nanostructured ceramics, it was concluded that PLS for broadband fluorescence spectra without variable selection could give a smaller error in temperature calibration than the method of measuring temperature by fluorescence intensities ratio from two temperature-coupled levels of activator energy [8]. Current paper dwells upon the possibility of further temperature calibration accuracy improvement when using interval PLS [9] with variable selection in combination with two more methods of multivariate data analysis – the principal component analysis (PCA) [10] and the hierarchical cluster analysis (HCA).

2 Experimental

The first sample to study was $98MgCaSrBaYAl_2F_{14}$-$2Ba(PO_3)_2$ glass with 0.5 mol% ErF_3. Excitation was carried out by CW laser diode at 980 nm and an optical power of 2.25 W. The spectra were registered in the 450-893 nm range with a resolution of about 0.3 nm and an error in the intensity less than 0.5 %. To conduct multivariate analysis, the measurements at 250 wavelengths were used in the range of 510-580 nm, covering the green bands of erbium ion fluorescence. The temperature of the sample varied from 94 to 508 K and was controlled with an accuracy 0.5 K.

Yb^{3+}:CaF_2, excited by a 1W CW laser diode with a spectral maximum at 967 nm, was the second sample investigated. The registration was carried out with a resolution of about 0.2 nm. For the multivariate analysis, 1024 spectral wavelengths were used in the 893-1107 nm range. The temperature of the sample varied from 339 to 423 K in 2 K increments and was controlled with an accuracy of not worse than 0.1 K.

3 Multivariate Analysis of Fluorescence Spectra

3.1 Outliers Detection by PCA

The description of the consecutive multivariate analysis proposed will be given on the example of temperature dependence of Er-doped 98MgCaSrBaYAl$_2$F$_{14}$-2Ba(PO$_3$)$_2$ glass fluorescence spectrum. The first stage of the multivariate analysis was PCA, which was carried out to detect outliers in order to get rid of the influence of unstable characteristics of the diode excitation and other uncontrolled factors. In most cases of multivariate analysis, PCA is the first method applied for the data dimension reduction, outlier detection, classification and regression.

In the case under analysis, the outlier detection was carried out by considering the temperature dependence of scores to the first principal component. For a certain temperature the considerable difference between score to the first principal component and the polynomial approximation is an explicit indicator that the result of the measurement at this temperature is an outlier and should be removed from consideration. Since the first principal component after the outlier elimination explains 0.997 of the total explained data dispersion for Er-doped 98MgCaSrBaYAl$_2$F$_{14}$-2Ba(PO$_3$)$_2$ glass, higher principal components inclusions are not appropriate. The outliers having been removed, 53 fluorescence spectra recorded at temperatures from 108 to 508 K remained for further analysis.

3.2 Forming a Training Dataset Using HCA

Thereupon, PLS must be applied to the remaining fluorescence spectra. PLS is the bi-linear statistical method, which operates on the predictor matrix $\boldsymbol{X}$ (in our case this is a matrix of fluorescence spectra with dimensions of 53 × 250) and the response vector $\boldsymbol{Y}$ (temperature vector with dimensions of 53 × 1). PLS is aimed at finding a small-dimensional subspace of the predictors and response, in which the covariance between the $\boldsymbol{X}$ and $\boldsymbol{Y}$ projections is maximal, whereas the calibration error is, therefore, minimal. The coordinate axes of the required subspace are latent structures, and their number is one of the factors determining the calibration model quality. Another factor affecting the calibration quality is the carrying out of a sufficient number of measurements to collect a training dataset in order to construct a model and to perform the cross-validation and a test dataset to execute the validation. Calibration accuracy can be described by two values of root-mean-square error. There are the root-mean-square error of cross-validation (RMSECV) in a training dataset and the root-mean-square error of prediction (RMSEP) in a test dataset:

$$\mathrm{RMSECV} = \sqrt{\sum_{training}\left(T - T_{predicted}\right)^2 \Big/ n_{training}} \quad \text{and} \quad \mathrm{RMSEP} = \sqrt{\sum_{test}\left(T - T_{predicted}\right)^2 \Big/ n_{test}} \qquad (1)$$

Here T and $T_{predicted}$ are the measured and the predicted values of temperature, $n_{training}$ and n_{test} are the numbers of measurements in the training and test datasets, respectively.

The 53 measurements available at our disposal will be divided into non-intersecting training and test datasets of the temperature values and the corresponding spectra. The optimal number of latent structures will be simultaneously determined, using the value of RMSEP.

When forming training datasets, uniform coverage of the range of response changes is often used. A good alternative in terms of reducing RMSEP is the Kennard-Stone algorithm [11]. According to this algorithm, the sample closest to the center of the response range is selected first, and each subsequent training dataset sample should be the most remote according to the metric used in the response space from those already selected. The best alternative to the above algorithms is the use of cluster analysis. Cluster analysis [13] is a method of multivariate data analysis that organizes objects into groups according to certain similar features. In the case under consideration, cluster analysis allows us to divide all the registered fluorescence spectra into a predetermined number of non-overlapping clusters. The training dataset includes one spectrum from each cluster. The minimum calibration error among the cluster analysis methods considered [12] was obtained by applying a hierarchical cluster analysis gradually combining the nearest fluorescence spectra or the already formed clusters to larger clusters. It should be noted that such an approach enables proceeding from the forming the training dataset in accordance with the values of the response to the analysis of spectra.

In the case under study, hierarchical cluster analysis is performed in the space of the first principal component of the fluorescence spectra. Under the additional condition of mandatory selection of two measurements corresponding to the extreme values of the scores to the first principal component, the training dataset contains from n to n+2 spectra in the allocation of n clusters. In [14] the measurements were chosen into the training dataset that were the most distant from the centers of clusters in order to take the diversity of data into account. In our case, the measurements closest to the centers of the clusters are chosen. In the one-dimensional principal component space, the approach demonstrated in [14] will inevitably worsen the quality of calibration due to the choice of spectra having adjacent values of scores to the first principal component.

Figure 1 shows the temperature dependence of the scores to the first principal component of the fluorescence spectra. It also depicts 8 hierarchically identified clusters into which the entire set of spectra is divided. The training dataset included 9 measurements ($n_{training}$=9 and n_{test}=44), indicated by black circles in Figure 1, since the maximum score value to the first principal component corresponds to the measurement closest to the center of the first cluster. The minimum value of RMSEP = 6.10 K in the test dataset is obtained to the projection to 5 latent structures, the corresponding value of RMSECV in the training dataset is 1.14 K. The quality criterion of the model is the value of RMSEP, since RMSECV decreases monotonically with increasing number of latent structures up to $n_{training}$–1.

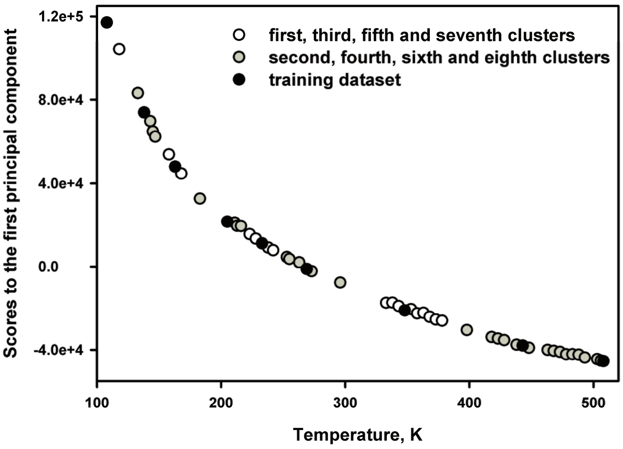

Figure 1. Dependence of scores to the first principal component on temperature for fluorescence spectra of a 0.5 mol% ErF_3 doped glass of $98MgCaSrBaYAl_2F_{14}$-$2Ba(PO_3)_2$, which was used to perform a hierarchical cluster analysis to determine the training dataset for temperature calibration by the PLS method.

3.3 Variable Selection by Searching Combination Moving Window Interval Projection to Latent Structures

After simultaneous separation of the matrix $\boldsymbol{X}$ and the vector $\boldsymbol{Y}$ into the training and test datasets and determining the number of latent structures for minimizing the RMSEP value for the full spectral range, it is crucial to select the variables or to optimize the spectral intervals in order to improve the quality of PLS modelling. A large number of optimization methods are characterized by the presence of the notions "interval" or "window" in their description [9,15-17]. In multivariate methods, an interval means either a single wavelength or several adjacent ones with correlated measurements, because they depend on the same physical or chemical parameters of the investigated objects. It is important that with a decrease in the number of wavelengths taken into account in modelling, the contribution of each of them increases. This can lead not only to an increase in calibration accuracy, but also to deterioration of the model stability due to the decrease in the measurements redundancy. Therefore, hereafter we will use not separate wavelengths, but spectral intervals, as is specific for spectroscopy.

In the simplest implementation of interval PLS (iPLS) [9], the full spectrum range is divided by a specified number of non-overlapping intervals with a separate model constructed for each of them. In order to achieve the minimum value of RMSEP (RMSECV), the spectral intervals can consistently merge (forward iPLS – fiPLS) or be extracted one by one from the full spectrum range (backward iPLS – biPLS) [15]. The interval that can change the size and can shift in the spectrum is usually called "window" [16, 17] in moving window iPLS method (mwiPLS) [17] and in searching combination moving window iPLS method (scmwiPLS) [17].

The developed modification of the scmwiPLS method [18] can be described as follows. When constructing PLS model, the minimum width of the spectral window should exceeds the number of latent structures by 1. In case considered of projection to 5 latent structures the minimum width of the spectral window contains 6 wavelengths. Similar to the iPLS method the position of the first such window is specified and fixed. The next window shifts within the entire spectral range of the measurements and merge with the first one. The minimum value of RMSEP specifies the position of the second spectral window. The procedure for increasing the number of windows continues similarly until the full spectral range is covered. So the global minimum of RMSEP can be found and the wavelengths are determined that should be included in the multivariate model.

3.4 Results and Discussion

Figure 2 shows the dependence of the minimum RMSEP in the test dataset using the scmwiPLS method on the number of spectral windows containing 6 wavelengths. The global minimum of the root-mean-square error of prediction $\mathrm{RMSEP}_{\mathrm{min}}$ = 4.42 K is presented by a filled square and corresponds to 9 windows or 54 wavelengths. Compared to PLS over the entire spectral range of fluorescence measurements that is presented by a filled circle, RMSEP for optimum variable selection by scmwiPLS method has decreased by 27%. The position of these 9 windows and the order of their selection in the multivariate model are demonstrated in Figure 3.

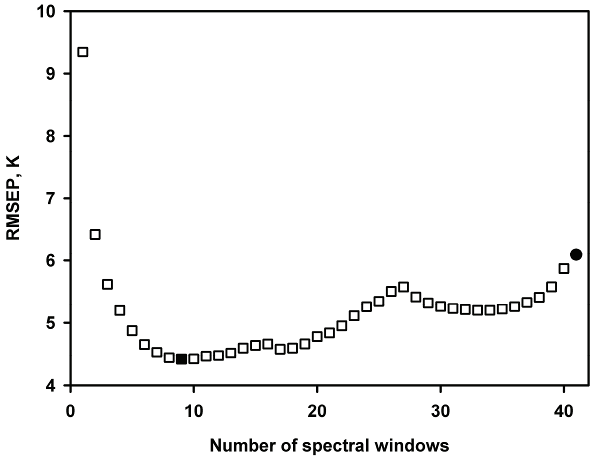

Figure 2. Dependence of RMSEP in the scmwiPLS method for fluorescence spectra of a 0.5 mol% ErF_3 doped glass of $98MgCaSrBaYAl_2F_{14}$-$2Ba(PO_3)_2$ on the number of spectral windows included in the model.

One can see that the temperature changes in the fluorescence intensity in the long-wave green band (to which the windows under numbers 1-4 and 7-9 refer) are more informative compared to the change in the short-wave band (windows 5 and 6). The centers of the first three selected windows are located near the wavelengths 537.3 nm, 547.2 nm and 559.5 nm, which correspond to the shortwave slope, the center and the long-wave edge of the band. In the windows selected, the fluorescence spectrum undergoes the fastest damping, so the multivariate simulation taking into account these intervals appears to be the most reasonable from the point of view of the physics of the phenomenon under consideration.

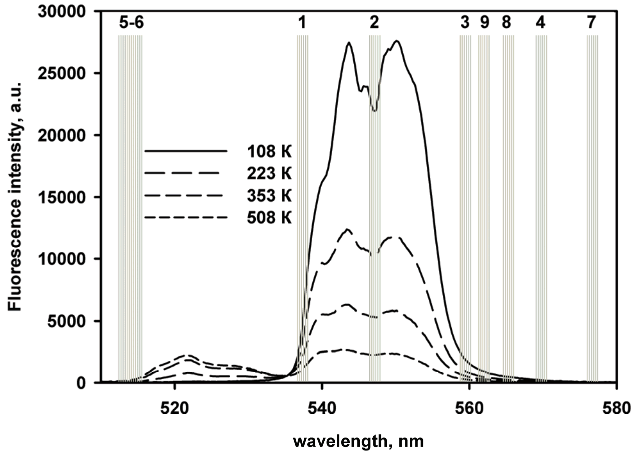

Figure 3. Fluorescence spectra of Er-doped glass considered and spectral intervals, for which the temperature prediction by the scmwiPLS method is characterized by a minimum value of $RMSEP_{min}$. The numbers above the intervals show the order of their selection in the model.

As for the second sample studied, the successive application of PCA, HCA and scmwiPLS to fluorescence spectra of Yb^{3+}:CaF_2 made it possible to determine the position of the spectral windows shown in Figure 4 and containing 136 wavelengths from 1024 ones in the measured fluorescence spectra, for which the projection to three latent structures provides RMSEP = 0.45 K. Compared to PLS over the entire spectral range of measurements (RMSEP = 0.93 K [18]), the root-mean square error of temperature prediction decreased more than twofold. It can be seen that for both cases analysed, the most informative parts of the investigated spectra are the regions of the most rapid change in fluorescence intensity depending on the temperature.

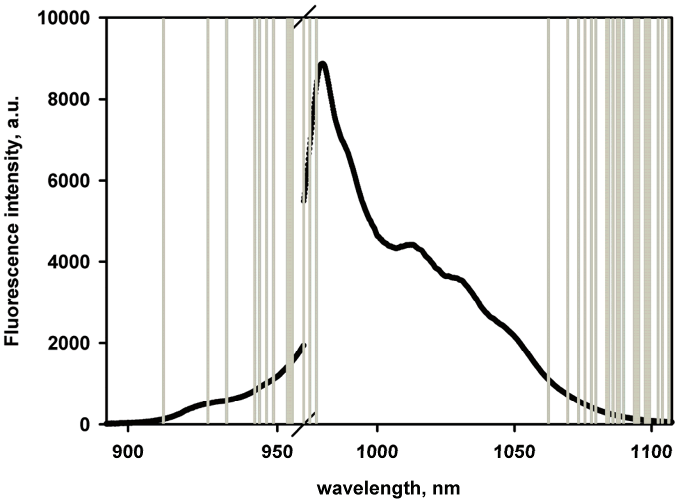

Figure 4. Fluorescence spectrum of a Yb^{3+}:CaF_2 at the temperature of 339 K and spectral intervals, for which the temperature prediction by the scmwiPLS method is characterized by a minimum value of RMSEP.

4 Conclusion

Examining the temperature dependence of the fluorescence of Er-doped $98MgCaSrBaYAl_2F_{14}$-$2Ba(PO_3)_2$ and Yb-doped CaF_2 and combined using multivariate methods for spectra processing showed the possibility of improving the accuracy of temperature calibration. Application of principal component analysis for outliers detection, hierarchical cluster analysis for forming training and test datasets, and searching combination of moving windows interval projection to latent structures for variable selection and temperature calibration makes it possible to significantly reduce the root-mean-square error of temperature calibration compared with the results obtained using only the ordinary projection to latent structures.

References

1. A.N. Babkina, M.A. Khodasevich, P.S. Shirshnev, “Temperature measurements using a projection to latent structures of fluorescence spectra of potassium–aluminum borate glasses with copper-containing molecular clusters,” *Optics and Spectroscopy*, vol. 122, no. 2, pp. 214-228, 2017.
2. A. Siai, P. Haro-González, K. Horchani Naifer, M. Férid, “Optical temperature sensing of Er^{3+}/Yb^{3+} doped $LaGdO_3$ based on fluorescence intensity ratio and lifetime thermometry,” *Optical Materials*, vol. 76, pp. 34–41, 2018.
3. D. Gong, T. Cao, Sh. Han, X. Zhu, A. Iqbal, W. Liu, W. Qin, H. Guo, “Fluorescence enhancement thermoresponsive polymer luminescent sensors based on BODIPY for intracellular temperature,” *Sensors and Actuators B: Chemical*, vol. 252, pp. 577– 583, 2017.
4. W.A. Pisarski, J. Pisarska, R. Lisiecki, W. Ryba-Romanowski, “Er^{3+}/Yb^{3+} co-doped lead germanate glasses for up-conversion luminescence temperature sensors,” *Sensors and Actuators A: Physical*, vol. 252, pp. 54–58, 2016.
5. G. Chen, R. Lei, F. Huang, H. Wang, Sh. Zhao, Sh. Xu, “Optical temperature sensing behavior of $Er^{3+}/Yb^{3+}/Tm^{3+}$:Y_2O_3 nanoparticles based on thermally and non-thermally coupled levels,” *Optics Communications*, vol. 407, pp. 57–62, 2018.
6. V.A. Aseev, Y.A. Varaksa, E.V. Kolobkova, G.V. Sinitsyn, M.A. Khodasevich, “Application of projection on latent structures for determining temperature of erbium-doped lead fluoride nano-glass-ceramics from upconversion fluorescence spectra,” *Optics and Spectroscopy*, vol. 118(5), pp.727-728, 2015.
7. H. Abdi, L.J.Williams, “Principal Component Analysis,” *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 2, pp. 433–459, 2010.
8. V.A. Aseev, Y.A. Varaksa, E.V. Kolobkova, G.V. Sinitsyn, M.A. Khodasevich, A.S. Yasukevich, “Comparison of two temperature measurement methods by upconversion fluorescence spectra of erbium-doped lead-fluoride nano-glassceramics,” *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, vol.15, no. 3, pp. 457–462, 2015.
9. L. Norgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, S. B. Engelsen, “Interval Partial Least-Squares Regression (iPLS): A Comparative Chemometric Study with an Example from Near-Infrared Spectroscopy,” *Applied Spectroscopy*, vol. 54, no 3, pp. 413–419, 2000.
10. K.H. Esbensen, P. Geladi, “Principal Component Analysis: Concept, Geometrical Interpretation, Mathematical Background, Algorithms, History, Practice,” *Comprehensive Chemometrics*, vol. 2, pp. 211–226, 2009.
11. R.W. Kennard, L.A. Stone, “Computer aided design of experiments,” *Technometrics*, vol. 11, pp. 137-148, 1969.
12. M.A. Khodasevich, N.A. Saskevich, “Training subset selection methods for calibration with fluorescence spectroscopy in small data sets of samples,” *Proceedings of the National Academy of Sciences of Belarus. Physics and Mathematics series*, vol. 54, no. 1, pp. 77–83, 2018 [in Russian].
13. I.D. Mandel, “Cluster analysis”, *Finance and statistics.* [in Russian: Klasternii analiz. Moskva. Financi i statistika], Moscow, 1988, pp. 176.
14. M. Daszykowski, B. Walczak, D.L. Massart, “Representative subset selection,” *Analytica Chimica Acta*, vol. 468, pp. 91-103, 2002.
15. X. Zou, J. Zhao, Y. Li, “Selection of the efficient wavelength regions in FT-NIR spectroscopy for determination of SSC of ‘Fuji’ apple based on BiPLS and FiPLS models”, *Vibr. Spectr*, vol. 44, pp. 220-227, 2007.
16. J.-H. Jiang, R. J. Berry, H. W. Siesler, and Y. Ozaki, “Wavelength Interval Selection in Multicomponent Spectral Analysis by Moving Window PLS Regression with Applications to Mid-Infrared and Near-Infrared Spectroscopic Data”, *Anal. Chem*, vol. 74, pp. 3555-3565, 2002.
17. Y. P. Du, Y. Z. Liang, J. H. Jiang, R. J. Berry, Y. Ozaki, “Spectral regions selection to improve prediction ability of PLS models by changeable size moving window PLS and searching combination moving window PLS “, *Anal. Chim. Acta*, vol. 501, pp. 183-191, 2004.
18. M.A. Khodasevich, V.A. Aseev, “Choice of spectral variables and increasing the accuracy of temperature calibration by projection on latent structures from the fluorescence spectra of Yb3+: CaF2,” *Optics and spectroscopy*, vol. 124 (5), pp. 713-717, 2018.