Electronic Supplementary Material (ESI) for Journal of Analytical Atomic Spectrometry.

Chemometric sorting based on laser-induced plume fluorescence: Characterization of spectral noise for effective preprocessing – Electronic supplementary information

Nai-Ho Cheung

Department of Physics, Hong Kong Baptist University, Kowloon Tong, Hong Kong, China

Abstract

The published article reports the spectral preprocessing of laser-induced plume fluorescence spectra and its denoising effect. Some details of the data treatment that are supplementary to the main arguments are presented here. They include (1) the step-by-step outlier rejection protocol, (2) a comparison of two outlier filters, (3) the evaluation of the noise due to $R_1$and R_2, (4) the evaluation of the noise counts due to the random C offset, and (5) the suppression of intra-class variance in favor of inter-class variance. Illustrative figures not directly relevant to the main theme are also presented here. They include the loading spectra used in the principle-component-analysis of red seal inks and Chinese black inks.

List of Figures and Tables

Outlier rejection protocol

As explained in the article, the dimmest spectra are outliers and have to be rejected. Conceptually, for a given set of N-MS spectra, say the segment centering on 500 nm, we expect the RSD of the Cr 520.6 nm peak height to drop when we do *n*-shot averaging. If the data distribute normally, the RSD should drop as $1/\sqrt{n}$. If there are outliers in the data set, the RSD would drop faster than $1/\sqrt{n}$. Based on this understanding, we design an outlier rejection protocol as follows.

1. We take the data set and do 1, 2, and 4 -shot averaging and plot RSD against $1/\sqrt{n}$. If the trend decays faster than linear, we know there are outliers.
2. We then rank the raw spectra by the 520.6 nm line brightness, from dimmest to brightest. We eliminate the dimmest *N* % from the set, regenerate the N-MS data, and repeat step 1. We keep increasing *N* until the plot of RSD against $1/\sqrt{n}$ is close to linear.
3. As an example, we list in Table E-1 the results for C5 red seal ink. The %RSD refers to the Cr I 520.6 nm line.
4. The plot of RSD *vs* $1/\sqrt{n}$ is shown in Fig. E-1, where four E*N* families with *N* = 0, 2, 8 and 12 are plotted.
5. As can be seen, the trendline becomes more linear with increasing *N*, indicating that the data set approaches a normal distribution and aberrations due to outliers are reduced.
6. We consider the E8 trendline to be adequately linear (exponent is 1 when rounded to one significant figure). We found that E8 corresponds to the rejection of raw spectra dimmer than about 10% of the average brightness. For example, the 16th dimmest spectrum has a 520.6 nm raw peak height of 2,744 counts while the average raw peak height among the 200 observations is 25,565 counts.
7. So, as a rule of thumb, raw spectra dimmer than 10% of the average brightness should be rejected.

Table E-1. %RSD of 520.6 nm line of C5 red seal ink at various stage of preprocessing.

	E0	E2	E8	E12
N-MS-1	22.9	18.6	6.51	5.33
N-MS-2	10.9	3.76	3.56	3.56
N-MS-4	3.17	2.88	2.53	2.38

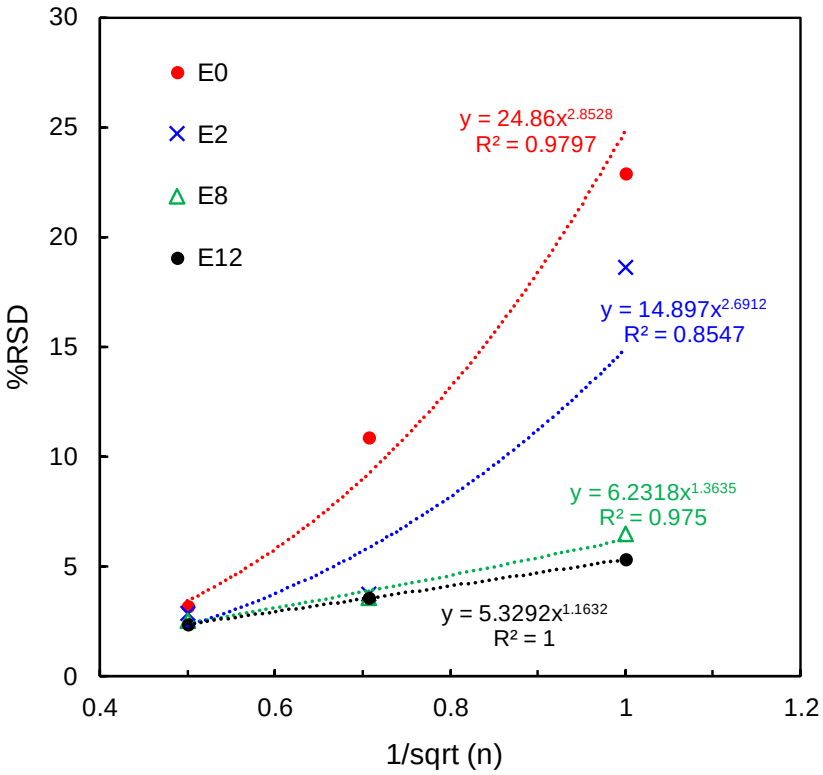

Fig. E-1. Effect of outlier rejection on %RSD of the Cr I 520.6 nm line height for C5 red seal ink N-MS spectra. The graph shows %RSD *vs* $1/\sqrt{n}$ for four rejected fractions: 0, 2, 8, and 12 %, respectively. Power-law trendlines and their associated R^2 are shown alongside.

Outlier filters based on (a) intensity of strong lines and (b) Hotelling T^2

We mentioned in the article that for minimally destructive PLIF analysis, sub-threshold ablation can occur. These events produce dim, blank spectra that should be rejected as outliers. In contrast, sub-threshold ablation seldom occurs in LIBS so outliers are not equated to the dimmest spectra. Instead, LIBS outliers are usually identified with the extremes in spectral brightness, both high and low.[1] Or more commonly, LIBS outliers are associated with observations far from the cluster center on a score plot, i.e., those with large Hotelling T^2.[2]

Here, we will use the red seal ink data set to show that dim PLIF spectra are not the same as T^2 outliers. As an example, we show in Fig. E-2 the PCA score plot of 200 observations of C3 ink for the λ segment centering on 500 nm. For this λ segment, the brightest feature is the Cr 520.6 nm line and it is maximally loaded for PC1 (see Fig. E-5). Low PC1 therefore correlates with dim (weak 520.6 nm line) spectra. The bottom 8% among them are circled in black in Fig. E-2. Quite obviously, they are inside the 95% T^2 ellipse and are not considered outliers.

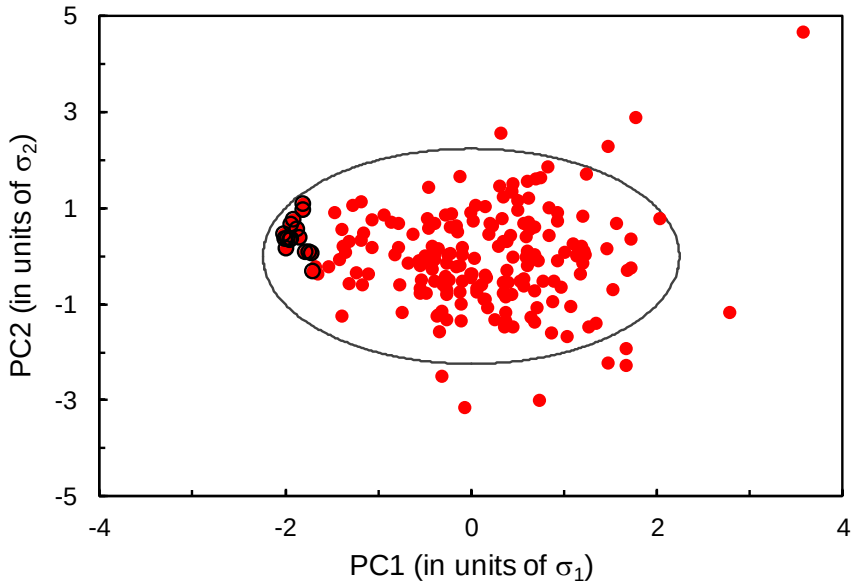

Fig. E-2. PCA score plot of C3 ink for the λ segment centering on 500 nm. The dimmest 8% are circled in black. The 95% Hotelling T^2 ellipse is also shown.

We will now show that the elimination of T^2 outliers will not help the sorting of seal inks. Fig. E-3 shows PCA score plots of seal inks preprocessed using N-MS-1-E8. The results of two outlier filters are compared. The left panel shows the rejection of the dimmest 8%. The right panel shows the rejection of the largest 8% T^2. Clearly, the T^2 filter does not help the sorting of seal inks.

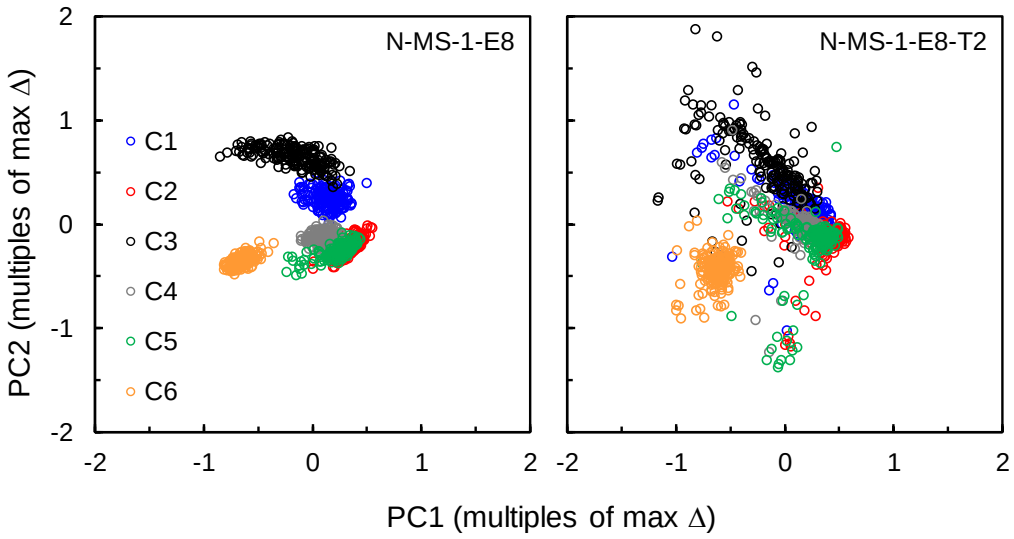

Fig. E-3. PCA score plots of red seal inks when two outlier filters are applied. Left: dimmest 8% eliminated. Right: largest 8% T^2 eliminated.

Evaluating the noise due to R_1 and R_2

Once the α noise is removed by area-normalization of the mean-subtracted spectrum, the contributions of the two noise ripples R_1 and R_2 can be estimated as follows. The N-MS-1-E8 spectrum $\hat{\tilde{E}}(\lambda)$ can be written as,

$$\hat{\tilde{E}}(\lambda) = \tilde{I}(\lambda)[1 + R_1(\lambda)] + \langle I\rangle R_1(\lambda) + \hat{R}_2(\lambda), \qquad \text{(E-1)}$$

where $\hat{R}_2(\lambda)$ is the normalized baseline ripple. At regions when $\tilde{I}(\lambda)$ is zero, the standard deviation $\hat{\sigma}(\lambda)$ of the $\hat{\tilde{E}}(\lambda)$ data set is due solely to $\langle I\rangle R_1 + \hat{R}_2$. Using C5 ink as an example, we examined $\hat{\sigma}(\lambda)$ around regions where $\hat{\tilde{E}}(\lambda) = 0$ and found that it was about 7. So, $\hat{\sigma}$ due to $\langle I\rangle R_1 + \hat{R}_2$ should be about 7, and [$\hat{\sigma}(\lambda)$ - 7] should be the noise spectrum due to R_1. We plot $\left|\hat{\tilde{E}}(\lambda)\right|$ (scaled by 0.065, red trace) and [$\hat{\sigma}(\lambda)$- 7] (blue trace) in Fig. E-4. They should track each other at the bright peaks when the $\langle I\rangle R_1 + \hat{R}_2$ ripples in Eq. (E-1) can be neglected. The overlap of the red and blue traces at the bright peaks is indeed evident from Fig. E-4. Two observations can be drawn. First, the 0.065 scale factor is consistent with the RSD of 6.51% of the N-MS-1-E8 entry in Table 1 of the article. It indicates the noise contribution due to R_1 at spectral peaks such as 520.6 nm. Second, at the dimmer regions, $\langle I\rangle R_1 + \hat{R}_2$ in Eq. (E-1) can no longer be neglected, so [$\hat{\sigma}$ (λ) – 7] will not track $\left|\hat{\tilde{E}}(\lambda)\right|$, as Fig. E-4 shows.

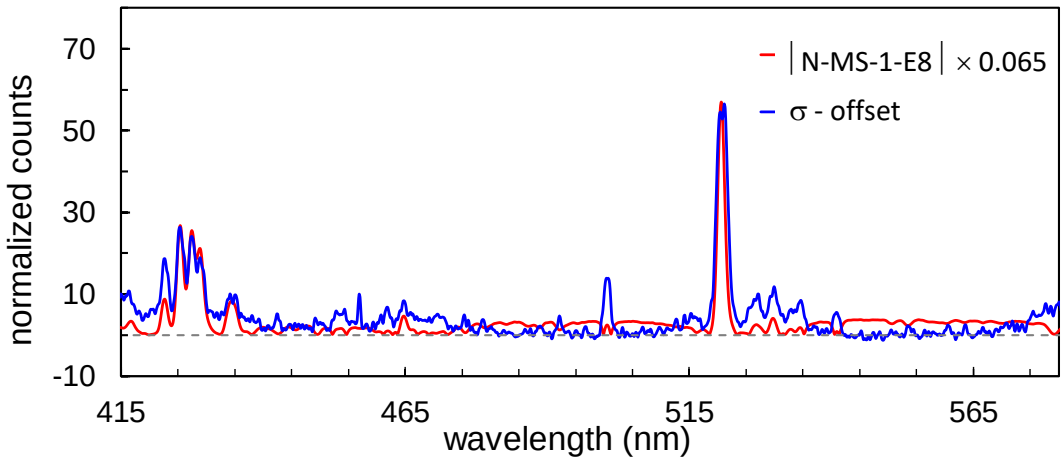

Fig. E-4. The average and the standard deviation of the C5 spectral dataset. Shown is the average of the | N-MS-1-E8 | spectra scaled down by 0.065, in red. The associated standard deviation minus an offset of 7 is shown in blue.

The noise counts due to $\langle I\rangle R_1$ and $\hat{R}_2$ (Eq. E-1) can be estimated like this. The noise counts due to $\langle I\rangle R_1 + \hat{R}_2$ is 7. The noise counts due to $\langle \tilde{I}\rangle R_1$, averaged over the spectrum, $\langle \hat{\sigma}\ (\lambda) - 7\rangle$, is about 4. $\langle\left|\tilde{I}\right|\rangle = 49$. $\langle I\rangle = 67$. So, noise counts due to $\langle I\rangle R_1$ = 5.5, and that due to $\hat{R}_2$ = 1.5.

Estimating the RSD due to the random offset *C*

We numerically simulated the PLIF spectra based on Eq. (1) of the article. We adjusted the various noise parameters to produce RSD that best match the empirical results shown in Table 1 of the article. We found that the noise counts due to *C* for the simulated C5 R-1 spectra was only 0.49 % of the 520.6 nm peak counts. But as pointed out in the article, the *C* noise counts should be compared against the mean spectral intensity rather than the 520.6 nm peak. The mean spectral intensity was 8% of the 520.6 nm peak, so *C* noise relative to it became a significant 6.2 %.

PCA loading spectra for sorting red seal inks

In the article, PCA of red seal inks was reported. Here, we show the PC loading spectra for the three preprocessed sets, N-MS-1-E0, N-1-E8, and N-MS-1-E8.

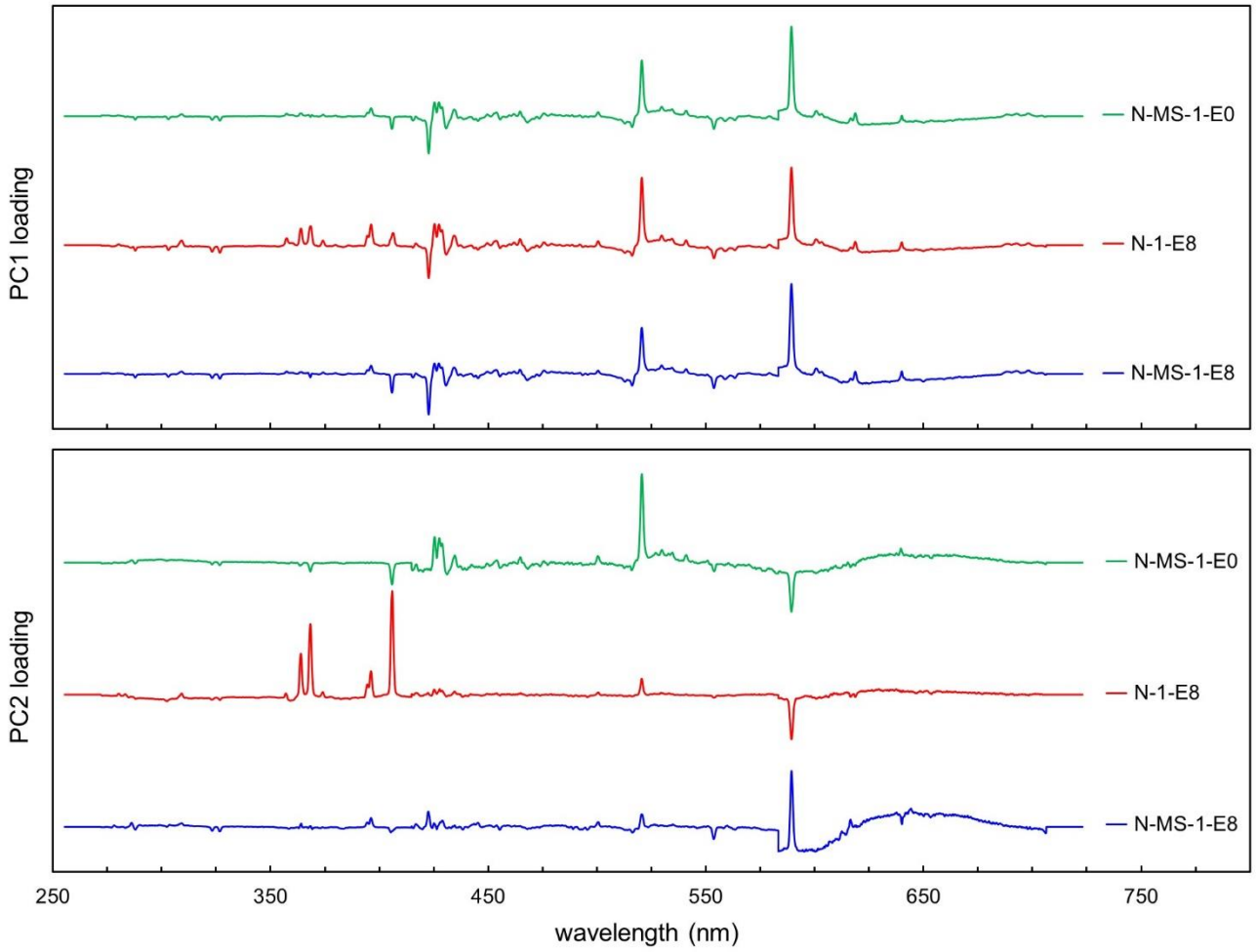

Fig. E-5. PC1 (upper panel) and PC2 (lower panel) loading spectra of the PCA of red seal inks based on three preprocessed data sets: N-MS-1-E0 (green trace), N-1-E8 (red trace), and N-MS-1-E8 (blue trace). The traces are offset vertically for clarity, and the leading and trailing pixels are zeroed to indicate the baseline.

Balancing inter and intra -class variance

In the article, we pointed out how the N-MS-1-E8 preprocessing suppressed the intra-class variances at 405.8 and 520.6 nm to favor the inter-class variances at the 589 nm shoulder. This is evident from the lower panel of Fig. E-5. This balancing of inter and intra -class variance can be quantified like this. We evaluate the SD at 520.6 nm and around the 589 nm shoulder (595.507 – 600.541 nm) for all six inks. Their ratios for various preprocessed data sets are tabulated below. We can correlate the SD ratio and the extent of cluster de-mixing (based on the score plots). Score plots for the bottom four rows of Table E-2 were given in Fig. 3 of the article. The score plot for the first row is shown in Fig. E-6. Based on these score plots, we can describe the cluster distribution as mixed or de-mixed, as shown in the third column of Table E-2. As can be seen from the table, ratio > 5 indicates mixing; ratio < 4 implies the Na shoulder gets weighted more than the Cr I 520.6 nm feature in the PC2 loading and the clusters are de-mixed along PC2.

Table E-2. Ratio of SD (520.6 nm) to SD (589 nm shoulder) for all six inks.

	σ(520.6)/σ(shoulder)	De-mixed?
N-2-E8	6.66	No
N-MS-1-E0	5.13	No
N-MS-1-E8	3.43	Yes
N-MS-2-E0	3.84	Yes
N-MS-2-E8	3.39	Yes

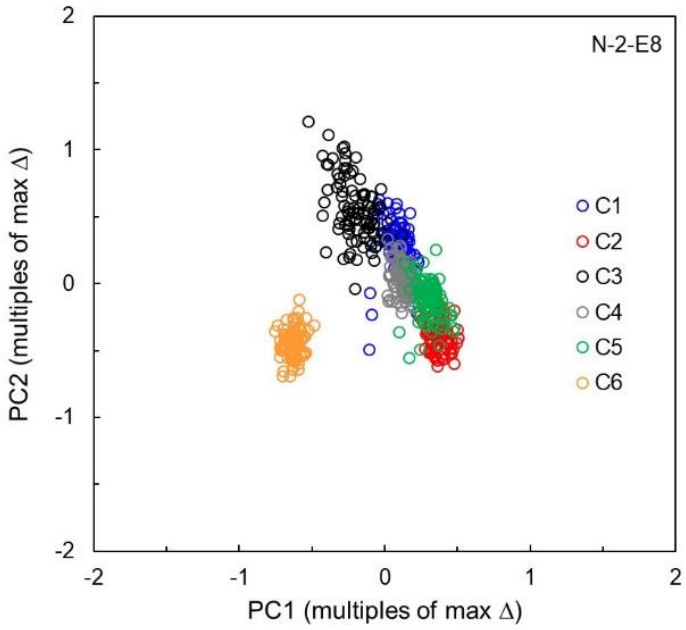

Fig. E-6. Score plot of PCA of red seal inks based on the N-2-E8 preprocessed data set.

PCA loading spectra for sorting Chinese black inks

In the article, PCA of Chinese black inks was reported. Here, we show the PC1 loading spectra for the two preprocessed sets, N-MS-1-E30 and N-1-E30.

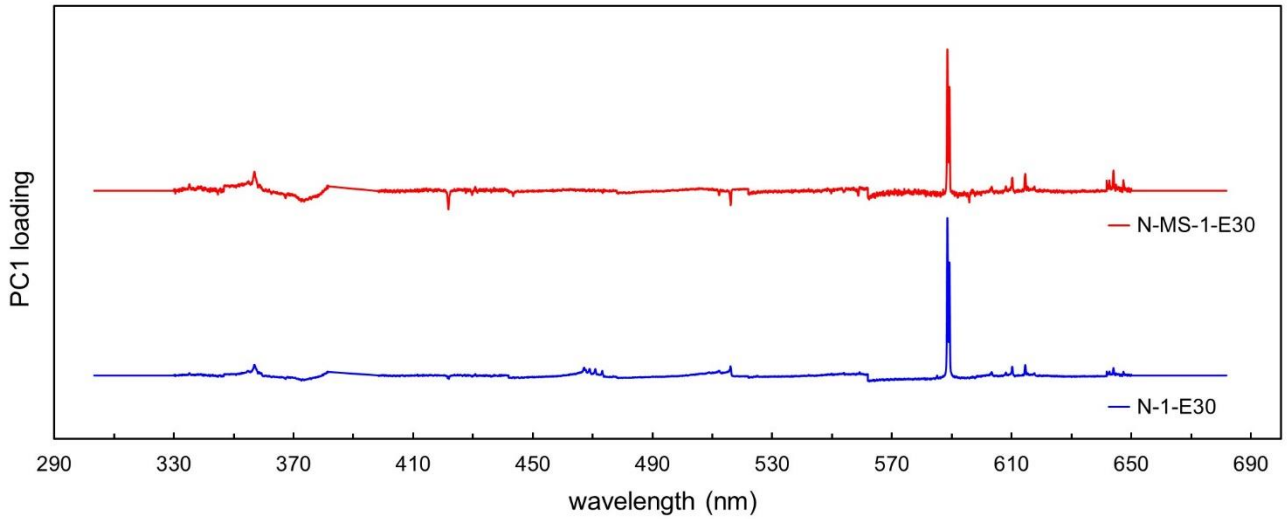

Figure E-7. PC1 loading spectra of the PCA of Chinese black inks on raw *xuan* based on two preprocessed data sets: N-MS-1-E30 (red trace) and N-1-E30 (blue trace). The traces are offset vertically for clarity, and the leading and trailing pixels are zeroed to indicate the baseline.

References

1 P. Porizka, J. Klus, D. Prochazka, *et. al., Spectrochim. Acta B,* 2016, **123,** 114-120.

2 L. Eriksson, T. Byrne, E. Johansson, J. Trygg, and C. Vikström, *Multivariate and Megavariate Data Analysis: Basic Principles and Applications,* 3rd Revised Ed., Umetrics Academy, Malmö, 2013.