Classification of Tea Quality by Two-Dimensional LED-Spectroscopy

Joel Gärtner Mikael Persson
jgartner@kth.se mikaepe@kth.se

Supervisor: Yongjiang Dong

JORCEP
The Sino-Swedish Joint Research Center of Photonics
Zhejiang University, Hangzhou, P.R. China

July 11, 2014

Abstract

We have worked on improving the one dimensional correlation analysis of fluorescence spectral data from dried Longjing tea leaves induced by illuminating the leaves with light emitting diods (LED). In addition we have investigated the possibilities of two dimensional correlation analysis. Results show no significant increase in performance for the two dimensional method.

1 Introduction

After plain water, tea is the most consumed beverage in the world (Awasom, 2011), thus any improvement in accuracy and/or efficiency on the classification of tea is of great interest. Previous work on classification of various kinds of food and beverage is extensive. Especially for the classification of tea, techniques used include (but is not limited to) electronic tongue (Palit et al., 2010), electronic nose (Chen et al., 2013), machine vision (Laddi et al., 2013), multi-spectral imaging (Wu et al., 2008) and laser induced spectroscopy (Mei et al., 2012).

While these techniques have their respective advantages, the simplicity, robustness and cheap price of LEDs makes them suitable for potential commercial applications. Work carried out by Dong et al. (2012) using LED induced fluorescence spectroscopy on the classification of tea has shown promising results. Present work is aimed to develop the LED fluorescence technique and investigate the possibilities of classification using two-dimensional correlation analysis on spectroscopic data from LED fluorescence.

2 Preparations and Experimental Setup

2.1 Preparations

We first spent some effort trying to automate as much of the data collection process as possible. Different kinds of electric motors controlled by software written in LabView2011 (National Instruments Corp., Austin, TX, USA) was tested and rejected due to poor performance. We also investigated the proper way of placing the LEDs and the tea samples to get high signal to noise ratios in the recorded spectrums.

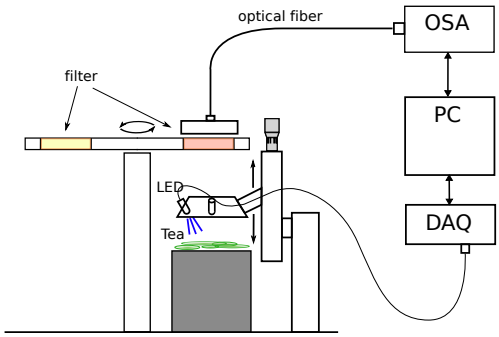

Figure 1: Schematic of the setup

2.2 Final Setup

The setup is shown in figure 1. It consists of four LEDs with wavelengths between 375 and 500 nm and a white light LED, placed above the the sample. Switching between the LEDs is done by manually arrange the connectors of a circuit board connected to a NI USB-6008 data acquisition device (DAQ). The DAQ is controlled by the LabView software to set the amplitude of the light. The sample was placed on black anodized aluminium with negligible fluorescence and the distance from LED to sample was manually controlled, see table 1 for the used distances. A long pass filter (Thor

D	Knob reading [mm]	Distance: Tea-to-LED [mm]
1	1	6
2	7	12
3	13	18
4	19	24
5	25	30

Table 1: Heights used during the measurements

Labs Inc, NJ, USA) of suitable wavelength is used to filter out reflectance from the LED. The filter cut versus the LED wavelengths is given in table 2. A multi-mode optical fiber connected to an Ocean Optics USB2000 spectrometer (Ocean Optics Inc., FL, USA) is used to record the spectrum data with the software written in LabView2011. The spectrometer integral time and LED-amplitude were fixed for the measurements with the single wavelength LEDs and the white LED respectively.

LED	λ [nm]	Filter cut [nm]
1	375	425
2	400	475
3	450	525
4	500	625
5	white	none

Table 2: LED vs. filter cut wavelength

3 Data analysis

3.1 One-Dimensional Analysis

For each group (specific LED and distance) we took two measurements for background noise without tea and six measurements for each tea grade. The average of the background spectrums were subtracted from the tea spectrums, which were then smoothed and normalized to the maximum value. Then followed the statistical analysis:

We first applied *Principal Component Analysis* (PCA), and a so called *elbow test* to determine the number of principal components to be used. Thereafter we used *Linear Discriminant Analysis* (LDA), to build a model for prediction of tea quality grading/pricing.

3.1.1 Principal Component Analysis

For large sets of data a Principal Component Analysis, PCA, allows you to select the important parts of the set. Mathematically, the PCA is a linear transformation into a new coordinate system arranged so that the first coordinate is aligned in the direction of greatest sample variance, the second coordinate is orthogonal to the first and in the direction of second greatest sample variance as possible and so on (Abdi and Williams, 2010).

Let us represent the data set of n samples of p-dimensional row vectors $\mathbf{x}_i^T$ (for example n samples of tea measured at p wavelengths) with the matrix $\mathbf{X}_{(n\times p)} = (\mathbf{x}_1 \cdots \mathbf{x}_n)^T$. Let us also assume $\mathbf{X}$ are zero-centered, i.e. the arithmetic mean of the components of each row vector $\mathbf{x}_i$ is zero. Finding the p-dimensional unit vector, referred to as *loading*, $\mathbf{w}_1$ in the direction of greatest variance is then

equivalent to maximizing the expression

$$\sum_{i}^{n}(\mathbf{x}_i \cdot \mathbf{w}_1)^2 = ||\mathbf{X}\mathbf{w}_1||^2 = \mathbf{w}_1^T\mathbf{X}^T\mathbf{X}\mathbf{w}_1 = \\ = \frac{\mathbf{w}_1^T\mathbf{X}^T\mathbf{X}\mathbf{w}_1}{\mathbf{w}_1^T\mathbf{w}_1} = R_{\mathbf{X}}(\mathbf{w}) \quad (1)$$

known as a *Rayleigh quotient*. For the symmetric matrix $\mathbf{X}^T\mathbf{X}$ a well known result now gives $\mathbf{w}_1$ as the normalized eigenvector corresponding to the greatest eigenvalue of $\mathbf{X}^T\mathbf{X}$. The first principal component score of $\mathbf{x}_i$ is now given as $P_{1i} = \mathbf{x}_i \cdot \mathbf{w}_1$. To find successive loadings the procedure is repeated on the matrix is given by subtracting the already found vectors from the previous matrix. The orthogonality of the eigenvectors of symmetric matrices ensures that the remaining principal components equals the remaining eigenvectors of $\mathbf{X}^T\mathbf{X}$. The full set of principal component scores can now be written as the matrix

$$\mathbf{P} = \mathbf{X}\mathbf{W} \quad (2)$$

where $\mathbf{W}_{(p\times p)}$ is the matrix of normalized eigenvectors of the matrix $\mathbf{X}^T\mathbf{X}$. Our data analysis is carried out using software written on Matlab2013a (The Mahtworks Inc, Natick, MA, USA).

3.1.2 Selection of principal components

Several methods exist to determine the number of principal components to be used (Karlis et al., 2003). For simplicity we choose to make our selection by an elbow test using a so called *scree plot*, see figure 3, of the eigenvalues of $\mathbf{X}^T\mathbf{X}$. By visual inspection of the plot, the components after the elbow is rejected. The number of components included in the analysis is denoted t.

3.1.3 Linear Discriminant Analysis

To build a predictive model for the quality of the tea, we remove one of the samples (i.e. the t number of PCs for the sample) from the set, one by one, and use the rest of the samples to build the model. This procedure is well proven for classification by spectroscopy (Mei et al., 2012; Dong et al., 2012)

If we have m variables to predict (in our case only one - the quality grade) and n samples, we let the predefined values (e.g. the grade given by experts or corresponding to a certain price level) form the matrix $\mathbf{Y}_{(n\times m)}$. The model is given by removing one of the rows from $\mathbf{Y}$ and solving the matrix equation

$$\mathbf{Y}_{(n-1\times m)} = \boldsymbol{\phi}_{(n-1\times t)}\boldsymbol{\theta}_{(t\times m)} \quad (3)$$

for $\boldsymbol{\theta}$ using linear regression. The *regressor* matrix $\boldsymbol{\phi}$ is given by

$$\phi_{kj} = \begin{cases} 1, & k = 0 \\ W_{kj}, & j = 1, 2, \ldots, t \end{cases} \qquad k \neq i \quad (4)$$

where the first column is to ensure that zero quality still has fluorescence spectrum data and $\mathbf{W}$ is the matrix of loadings, given above. The predicted values for the left-out sample is now given by:

$$[Y_{i1}, Y_{i2}, \ldots, Y_{im}] = [\phi_{i1}, \phi_{i2}, \ldots, \phi_{1t}]\boldsymbol{\theta} \quad (5)$$

The procedure is repeated for all of the samples to build the complete model. To describe the quality of the discrimination we define a classification index Q as

$$Q = \frac{|\mu_{Y_j\subset i} - \mu_{Y_j\not\subset i}|}{\sigma_{Y_j\subset i} + \sigma_{Y_j\not\subset i}} \quad (6)$$

where μ is the mean for the samples of class i and outside class i respectively. Similarly σ is the square root of the sample variation.

3.2 Two-Dimensional Analysis

A two-dimensional (2D) correlation analysis on spectroscopic data can provide additional information, not available from the one dimensional analysis. Essentially one can incorporate an additional variable in the analysis. The technique stems from two dimensional infrared spectroscopy, (Noda, 1990), and was subsequently generalized to incorporate many more fields of spectroscopy (Noda, 1993).

3.2.1 Generalized 2D spectroscopy

We define the *dynamic* spectra as $y(\nu, x)$ where x is the second variable (e.g. the height) varying from x_1 to x_2, i.e. the intervall X, and ν is the frequency. To obtain the 2D correlation spectra we *Fourier transform* the dynamic spectra from the x-domain to the frequency domain. If we measure

the dynamic spectral fluctuations at the frequency ν_1 the Fourier transform is given by

$$Y_1(\omega) = \int_{x\in\mathbb{R}} y(\nu_1, x)e^{-i\omega x}\mathrm{d}x = \\ = Y_1^{\mathrm{Re}}(\omega) + iY_1^{\mathrm{Im}}(\omega) \tag{7}$$

and the similarly the *conjugate* of the Fourier transform at ν_2:

$$Y_2^*(\omega) = \int_{x\in\mathbb{R}} y(\nu_2, x)e^{+i\omega x}\mathrm{d}x = \\ = Y_2^{\mathrm{Re}}(\omega) - iY_2^{\mathrm{Im}}(\omega) \tag{8}$$

We now define the complex two dimensional correlation intensity:

$$\Phi(\nu_1,\nu_2) + i\Psi(\nu_1,\nu_2) = \frac{1}{\pi X}\int_0^\infty Y_1(\omega)Y_2^*(\omega)\mathrm{d}\omega \tag{9}$$

The real part $\Phi(\nu_1, \nu_2)$ is referred to as the synchronous correlation spectrum which describes the similarity of the x-dependent behaviour at (ν_1, ν_2). The imaginary part $\Psi(\nu_1, \nu_2)$ is called the asynchronous correlation spectrum and describes the difference in x-dependent behaviour. These are commonly visualized using 2D contour maps.

3.2.2 Tea classification, application

When using 2D-correlation with the second variable x as time, one can use well established procedures to draw conclusions from the synchronous- and asynchronous maps, c.f. *Noda's rule* (Noda, 1990). We on the other hand use x as the distance from LED to sample or the LED wavelength. Therefore we decided to use the same procedure as above, namely PCA followed by LDA. We took the correlation intensities from our correlation maps as data points to form our matrix with rows for each of the six captures and seven tea samples. Due to the large amount of data we made the analysis on a smaller part of the maps.

4 Results

4.1 One dimensional analysis

Seven grades of Longjing teas were used. For each grade we recorded the spectrum six times for each of the five LED types and each of the five heights - in total 1050 spectrums. An example of the recorded spectrums for one LED and height is shown in figure 2.

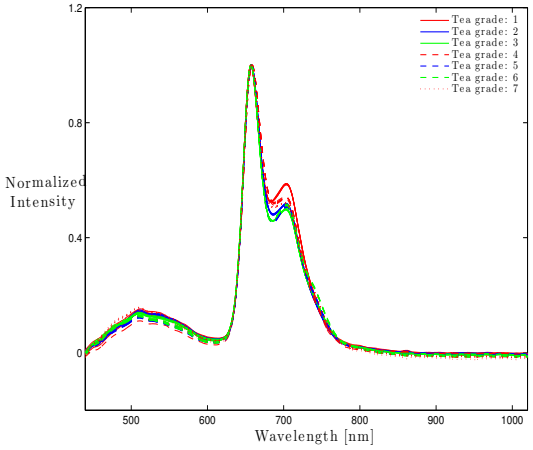

Figure 2: Example of the recorded spectrums, 6 for each of the seven tea grades. LED wavelength 375 nm, height number 3, i.e. 18 mm

For the same group of measurements (375 nm LED and heigth = 18 mm) the first three principal component loading vectors, the scree plot, the discrimation indices and the correlation is shown in figures 3 through 6. After performing the data analysis we compared the correlation coefficients of predicted price vs. price and predicted grade vs. grade respectively. We generally got fairly high correlation coefficients, above 0.92, see table 3.

	6 mm	12 mm	18 mm	24 mm	30 mm
375	0.972	0.982	0.976	0.973	0.982
400	0.963	0.971	0.979	0.986	0.993
450	0.947	0.969	0.970	0.978	0.961
500	0.967	0.973	0.974	0.966	0.966
white	0.973	0.982	0.974	0.949	0.966

Table 3: Correlation, predicted tea price and tea price for the five LED's and heights.

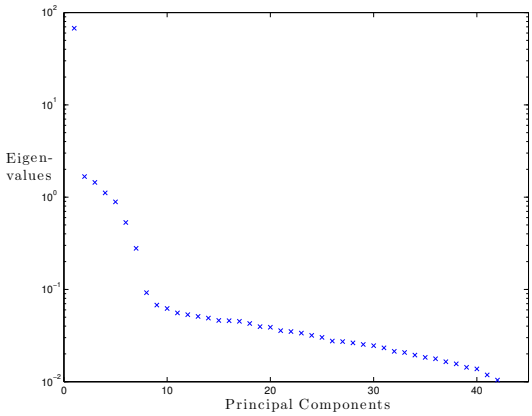

Figure 3: Scree plot: The elbow is identified at the tenth principal component.

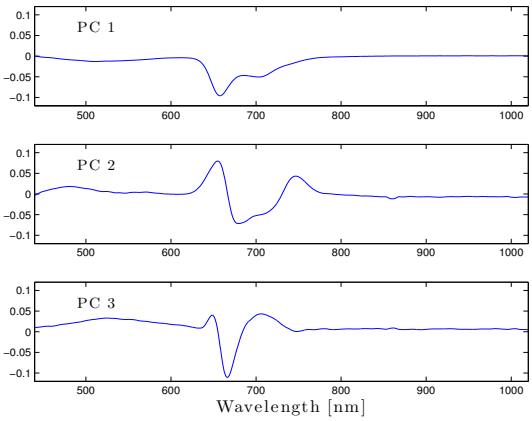

Figure 4: The first three PC loadings

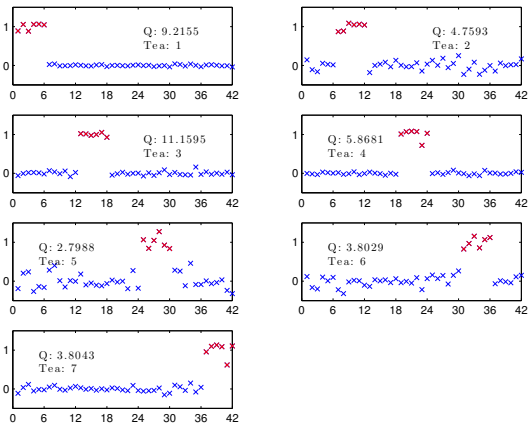

Figure 5: Example of the discrimination index Q.

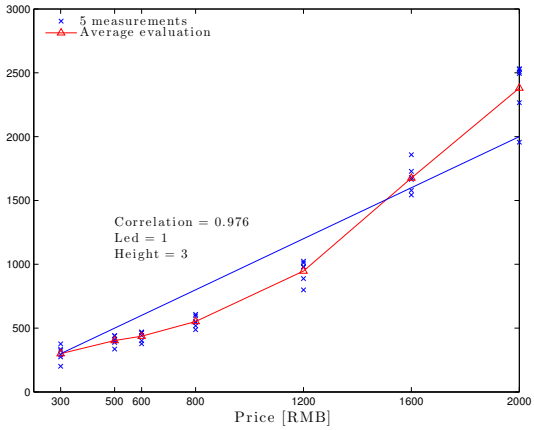

Figure 6: The correlation coefficient for the 375 nm LED and third height (18 mm).

We noted the following:

- We generally got better correlation coefficients for tea price than for tea grade, se table 4 for the price of the specific grades
- The 400 nm LED gave best average correlation coefficient, i.e. the average over the five five distances, as well as the best overall correlation coefficient, 0.993 for the fifth height, 30 mm
- The 375 nm LED gave the best minimum correlation coefficient

Similarly when analyzing the the different heights we noted the following:

- The second height (12 mm) gave the best average correlation coefficient, although the difference from the third and fourth height are negligable
- The third height (18 mm) gave the best minimum correlation coefficient
- The fifth height gave the best correlation coefficient for the 400 nm LED, c.f. above

We also added several heights together and several LEDs together respectively, to investigate if the correlation results could improve. The same type of 1D analysis were performed as above (PCA and LDA) on the new larger set of data but from this we could not see any significant change in the results. The best result was achieved when using heights 2,3 and 4 and LED 1 or 2.

Tea grade	Tea price [RMB]
1	300
2	500
3	600
4	800
5	1200
6	1600
7	2000

Table 4: Tea price for the seven tested grades

4.2 Two dimensional analysis

H	synch.	asynch.	LED	synch.	asynch.
375	0.976	0.981	1	0.963	0.972
400	0.976	0.975	2	0.972	0.970
450	0.978	0.974	3	0.978	0.979
500	0.972	0.950	4	0.976	0.976
W	0.960	0.946	5	0.978	0.985
Avg	0.972	0.965	Avg	0.973	0.976

Table 5: 2D correlation coefficients when using x as height and using x as wavelength

The result of two dimensional analysis did not significantly deviate from the one dimensional analysis. The correlation coefficients were more stable over the LEDs and heights respectively than for the 1D analysis, see table 5. Examples of principal component loading vectors are given in figures 7 and 8. See the appendix for examples of synchronous and asynchronous maps. The spread between the 6 measurements about their average predicted grade/price were lower for the two dimensional analysis.

5 Summary and Discussion

Building on previous work in one dimensional LED spectroscopy we have shown that the most effective wavelengths for the classification of dried tea leaves are in the lower visible span, about 400 nm. This is possibly due to the fact that some of the important active substances in the tea will emit in the region around 500 nm which will partially be filtered away if using longer wavelength LEDs. We have also shown that for very short LED-to-sample distances, signal to noise ratios were reducing thus

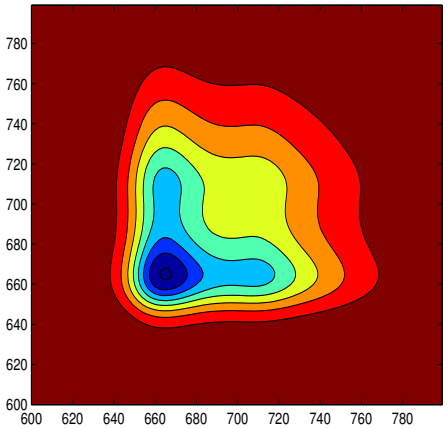

Figure 7: 400 nm LED: The first principal component loading vector for the synchronous maps as function of wavelength and wavelength [nm]. Only part of the dynamic spectrum is shown. The principal component loading is computed as a vector that consist of each row of the above plotted matrix. We choose to plot the matrix as a contour map for better visualisation.

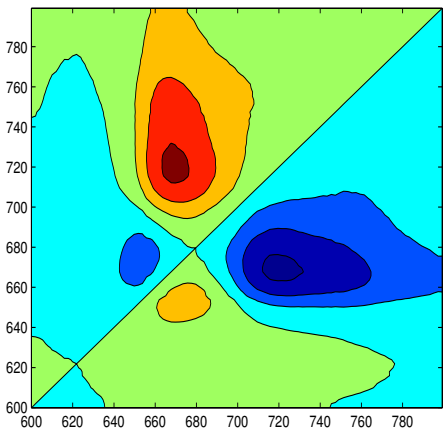

Figure 8: 400 nm LED: The first principal component loading vector for the asynchronous maps as function of wavelength and wavelength [nm]. Part of the dynamic spectrum is shown.

reducing the discrimination performance of the system. We believe this is due to the specific construction of the system, where the LED light comes out of vision for the optical fiber. The fifth height (30 mm) also show signal to noise ratios below desired.

In the 2D maps the tea samples were easier to distinguish visually from each other but the performance of the 2D method did not prove better than the 1D method. This was somewhat unexpected, perhaps there are better methods for analysing 2D data than PCR.

One possible reason for not achieving higher correlation coefficients could be that the tea quality is not fully reflected in the price and/or grade number as presently used. This could be the reason for the correlation plots to generally display a "concave" appearance, meaning the prediction is undershooting for the mid samples. It would be interesting to test correlation with tea expert grade or tasters' mark as well.

5.1 Ideas for improvements and future work

Generally, we would like to work on increasing the automation for the data collection process. For example, the problem with low signal to noise at the fifth height could possibly be solved by adjusting the integral time and/or the LED amplitude level. This was not practicable for our work. Also, one could increase the number of heights to further examine the possibilities of 2D analysis. Further reducing background noise is also desired, for example by conducting the data collection at night or enclosing the system to reduce ambient light.

Although one could come up with new ideas for the second variable in the 2D analysis (such as varying the sample temperature/humidity/air pressure, varying the sample size or the level of grinding, varying the solution concentration for liquid tea etc.), these methods seem inapplicable to future commercial uses where the device needs to be compact and relatively cheap. This makes future research using cheap LED's and relatively simple methods highly interesting.

6 Acknowledgements

First of all we would like to thank our supervisor mr. Yongjiang Dong for the time, effort and knowledge he put into supporting our work. We also thank the Zhejiang Provincial Key Laboratory for Sensing Technologies for providing equipment and material. Finally we thank all the members of JORCEP, the Sino Swedish Joint Research Center of Photonics, who made all the arrangements during our stay in Hangzhou.

References

H. Abdi and L. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews - Computational Statistics*, 2(4):433–459, 2010.

I. Awasom. Tea. *Journal of Agricultural and Food Information*, 12(1):12–22, 2011.

Q. Chen, A. Liu, J. Zhao, and Q. Ouyang. Classification of tea category using a portable electronic nose based on an odor imaging sensor array. *Journal of Pharmaceutical and Biomedical Analysis*, 84:77–83, 2013.

Y. Dong, X. Liu, L. Mei, C. Feng, C. Yan, and S. He. Led-induced fluorescence for tea quality assessment. Centre for Optical and Electromagnetic Research, Zhejiang Provincial Key Laboratory for Sensing Technologies, Zhejiang University, 2012.

D. Karlis, G. Saporta, and A. Spinakis. A simple rule for the selection of principal components. *Communications in Statistics, Theory and Methods*, 32(3):643–666, 2003.

A. Laddi, S. Sharma, A. Kumar, and P. Kapur. Classification of tea grains based upon image texture feature analysis under different illumination conditions. *Journal of Food Engineering*, 115:226–231, 2013.

L. Mei, P. Lundin, M. Brydegaard, S. Gong, D. Tang, G. Somesfalean, S. He, and S. Svanberg. Tea classification and quality assessment using laser-induced fluorescence. *Applied Optics*, 51(7):803–811, 2012.

I. Noda. Two-dimensional infrared (2d ir) spectroscopy: Theory and applications. *Applied Spectroscopy*, 44(4): 550–561, 1990.

I. Noda. Generalized two-dimensional correlation method applicable to infrared, raman and other types of spectroscopy. *Applied Spectroscopy*, 47(9):1329–1336, 1993.

M. Palit, B. Tudu, P. Dutta, A. Dutta, A. Jana, J. Roy, N. Bhattacharyya, R. Bandyopadhyay, and A. Chatterjee. Classification of black tea taste and correlation with tea taster's mark using voltammetric electronic tounge. *IEEE Transactions on Instrumentation and measurement*, 59 (8):2230–2239, 2010.

D. Wu, H. Yang, X. Chen, Y. He, and X. Li. Application of image texture for the sorting of tea categories using multispectral imaging technique and support vector machine. *Journal of Food Engineering*, 88:474–483, 2008.

Appendix: 2D maps

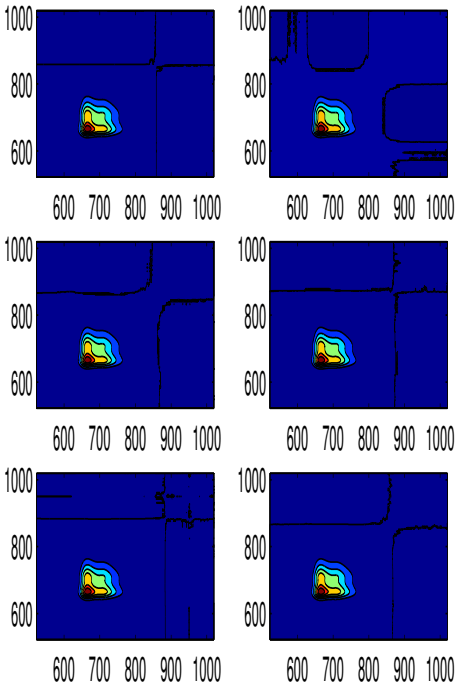

Figure 9: 400 nm LED: the six captures of the second tea grade and their synchronous maps.

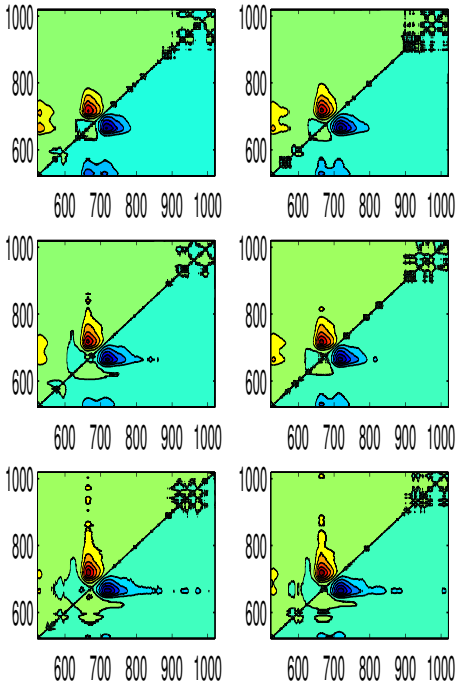

Figure 10: 400 nm LED: the six captures of the second tea grade and their asynchronous maps.

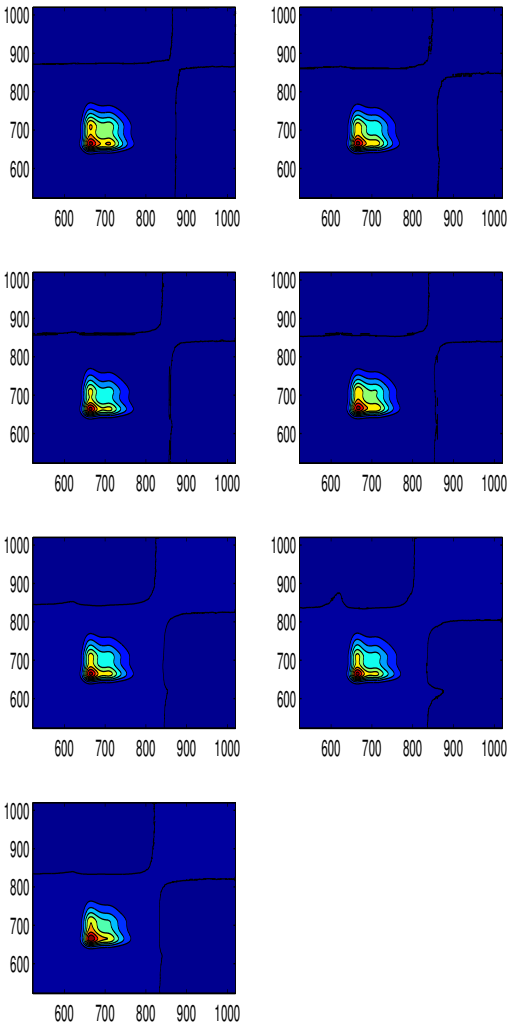

Figure 11: 400 nm LED, synchronous correlation maps for the seven tea samples.

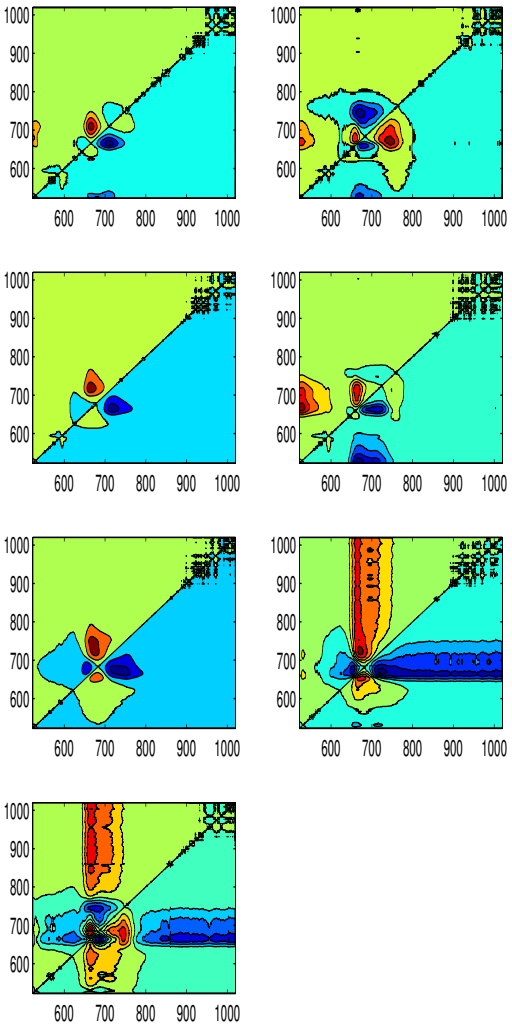

Figure 12: 400 nm LED, asynchronous maps for the seven tea samples.