The Korean Journal of Applied Statistics (2010)
23(1), 179–187

Clustering Red Wines Using a Miniature Spectrometer of Filter-Array with a Cypress RGB Light Source

Kyungmee Choi[1]

[1]College of Science and Technology, Hongik University

(Received January 2010; accepted January 2010)

Abstract

Miniature spectrometers can be applied for various purposes in wide areas. This paper shows how a well-made spectrometer on-a-chip of a low performance and low-cost filter-array can be used for recognizing types of red wine. Light spectra are processed through a filter-array of a spectrometer after they have passed through the wine in the cuvettes. Without recovering the original target spectrum, pattern recognition methods are introduced to detect the types of wine. A wavelength cross-correlation turns out to be a good distance metric among spectra because it captures their simultaneous movements and it is affine invariant. Consequently, a well-designed spectrometer is reliability in terms of its repeatability.

Keywords: Spectrometer, cross-correlation, pattern recognition.

1. Introduction

There have been more and more needs for a cheap and well-made miniature spectrometers not only because of their small, light-weight, and non-fragile properties, but also because of their variety of applications. Recent works (Morawski, 2006; Chang and Lee, 2008) have reviewed the applications of spectrometers which provide solutions to a variety of promising applications in biological, chemical, medical, or pharmaceutical industries, and also showed that the filter-based spectrometers have high resolution. There has also been an effort to recover the original target spectrum based on a linear model which has to confront an inversion of the huge matrix whose columns are highly correlated (Chang and Lee, 2008).

Because the miniature spectrometers are often cheap, filters are not exactly a delta function for a given narrow range response. Thus the response spectrum out of the miniature spectrometer of low-cost filter-array is distorted and augmented by the responses from other adjacent filters. However their multiple filters and detectors help avoid moving elements and capture the target spectrum in a very short time (Chang and Lee, 2008). So, their random variation is manageably small, which

This research was supported by the research fund of Hongik University in 2007.
[1]Professor, College of Science and Technology, Hongik University at Jochiwon, Sinan 300, Jochiwon, Yungi Chung-Nam 339-701, South Korea. E-mail: kmchoi@hongik.ac.kr

guarantees the repeatability of measurement. In other words, if a material is measured repeatably under the same condition, the miniature spectrometer provides very similar distorted curves. In this paper, digital signal processing(DSP) (Duda *et al.*, 2001; Peebles, 2000) techniques are applied to the distorted output spectra from the miniature spectrometer to show that it is reliable in terms of its repeatability. That is, this tiny chip can be used to recognize types of red wine even with those distorted curves without recovering the original target spectrum.

Four types of wine are prepared in the clear cuvettes, and a miniature spectrometer of low-cost filter-array measures output light spectra which have passed through the wine. As an unsupervised learning method, agglomerative hierarchical clustering(AHC) is used to discover the underlying relationship among output spectra. Here, as an analog of time cross-correlation, wavelength cross-correlation is defined to evaluate the simultaneous movement of the output spectra. A criterion pseudo-F (Hastie *et al.*, 2001; Rencher, 2002) is used to estimate the proper number of underlying clusters. Once the clusters are determined, an intuitive and simple way of classification method is introduced based on the cross-correlation. For the analysis, we do not consider any dimension reduction or feature extraction methods (Hasite *et al.*, 2001; Johnson and Wichern, 2007; Krzanowski, 2000; Rencher, 2002) such as principal component analysis(PCA), independent component analysis(ICA), reduced rank regression(RRR), or canonical correlation analysis(CCA) because these methods would corrupt the cross-correlation. The results are going to be interpreted in terms of repeatability of the spectrometers.

The suggested methods in this paper can be applied to analyze not only the data from the miniature spectrometer but also the ones from typical spectrometers. That is, the scope of the methodology can be extended from the wine detection to all liquid material detection, and further to clustering materials in a variety of application fields like biological, chemical, medical, or pharmaceutical industries.

Section 2 explains how the spectrometer has collected the output spectra passed through cuvettes which contain 4 different types of wine. Section 3 provides the wavelength cross-correlation as a distance metric and the image of correlation matrix. Section 4 discusses AHC as an unsupervised learning to find proper underlying groups of the data based on the correlation. In Section 5, a simple classification method is suggested and practiced to the wine data. Finally, section 6 draws the summary and conclusion.

2. Experiments

Twelve cuvettes are prepared to contain four different types of red wine, and each three cuvettes contain the same type of wine. There are two replications, so that there are in total twenty four output spectra. A cheap fine miniature spectrometer on-a-chip based on the 120 nano-optical filter-array measures output spectra sesitivity which have passed through the cuvettes of wine. This spectrometer converts the input optical signal into the output electrical signal in a few nanometer scale size without any moving parts. A surface plasmon technology based on metallic nanowires allows the integration for manufacturing nano-optical filters onto a regular CCD/CMOS imager production. Depending on the type of light source the emission power and the spectra can also vary significantly across devices and operating conditions. In this experiment, a cypress RGB light was used and its camera shutter speed was 15.5m/sec.

Figure 2.1 illustrates 6 output spectra for each 4 kinds wine data along with the 120 wavelengths corresponding to an array of filters. It is generally hard to tell by a sight whether the spectral

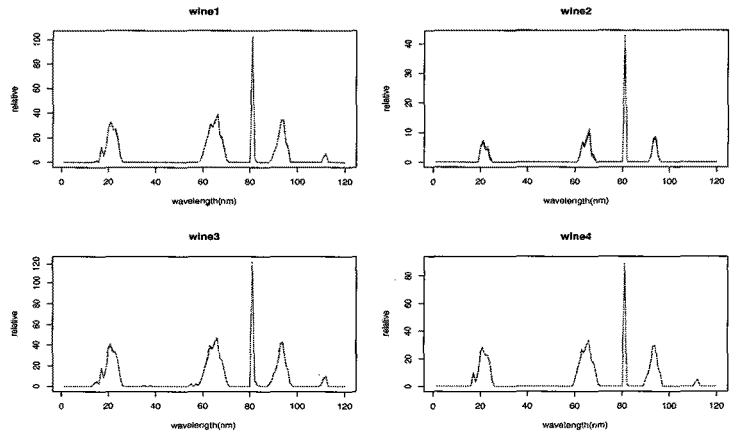

Figure 2.1. Plots of 6 output spectra from a spectrometer of 120 filter-array for 4 kinds of red wine.

signals from the spectrometer are the same kind or not. The four types of wine spectra are not also easily classifiable by a sight. So in order to pursue a systematic process of discriminating the four wine types, the statistical clustering methods are adopted.

Since the filters are not a delta function for a given narrow range of wavelength, the response spectrum out of the miniature spectrometer is distorted by each filter's spectral characteristics. However if the repeatability of measurement is guaranteed, digital signal processing(DSP) techniques can help control the random variation among the data and the spectrometer can cluster and classify the material types. Reversely, if the types of materials are well detected by DSP, the spectrometer can be thought to provide a reliable measurement based on their fine transmittance or reflection properties.

3. The Correlation Coefficient as a Metric

Euclidean distance has been often used to measure the distance between vectors in clustering analysis (Duda *et al.*, 2001; Hasite *et al.*, 2001; Johnson and Wichern, 2007; Krzanowski, 2000; Rencher, 2002). However typical distance metrics, like Euclidean distance, change when either one of the vectors is shifted in its location or scaled by a number. Contrarily, the correlation coefficient is known to be affine invariant, which means this statistic remains the same under both location and scale transformations of the data. The correlation coefficient has been applied to catch the exact simultaneous movement of two signals, or to study the stationarity of the signals over time in DSP. In this paper, we use the correlation coefficient (Box *et al.*, 1994; Peebles, 2000) as a metric to group spectra based on their synchronizing property.

For p sequential wavelength λ_i and observations $(x_{\lambda_1}, y_{\lambda 1}), (x_{\lambda_2}, y_{\lambda_2}), \ldots, (x_{\lambda_p}, y_{\lambda_p})$ of a pair of spectra (X, Y), the wavelength cross correlation can be defined by

$$r(X,Y;\lambda) = \frac{p\sum_{i=1}^{p} x_{\lambda_i} y_{\lambda_i} - \left(\sum_{i=1}^{p} x_{\lambda_i}\right)\left(\sum_{i=1}^{p} y_{\lambda_i}\right)}{\sqrt{p\sum_{i=1}^{p} x_{\lambda_i}^2 - \left(\sum_{i=1}^{p} x_{\lambda_i}\right)^2}\sqrt{p\sum_{i=1}^{p} y_{\lambda_i}^2 - \left(\sum_{i=1}^{p} y_{\lambda_i}\right)^2}}. \tag{3.1}$$

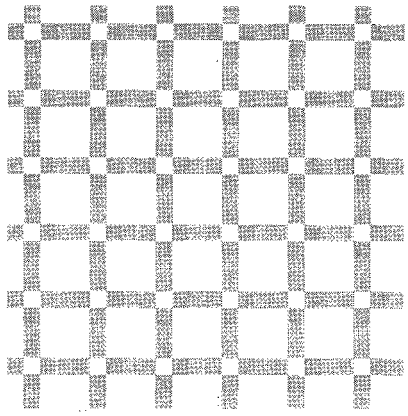

Figure 3.1. Image of the correlation matrix when the data are arranged periodically in the order of 4 wine types.

The correlation coefficient can range from -1 to 1. If it is $+1(-1)$, there is a perfect positive (negative) synchronization between two spectra. That is, for positive r, if X increases, so does Y. For negative r, if X increases, Y decreases. If it is 0, there is no synchronization between two spectra. Adopting this statistic to a spectrometer of filter-array, we can consider p as a necessary number of filters on it. In DSP, the time shift is used to express the the current process as a finite, linear aggregate of previous values of the process (Peebles, 2000). Here, not like in DSP, we do not consider the wavelength shift because the data has been collected under the well controlled environment, and we do consider the standardized form because of its affine invariant property.

For any real numbers a, b, c and d, we can easily prove its affine invariant property since

$$r(aX + b, cY + d; \lambda) = r(X, Y; \lambda). \tag{3.2}$$

In the real situations, the cuvettes will be replaced by other containers which could vary in their shape, thickness, or color. The spectra would change in both their location and scale. Still, the correlation coefficient could stay the same.

In order to look for any pattern in the correlation matrix, we arranged the data in a periodic order such as type 1, type 2, type 3, type 4, type 1, type 2, ... Figure 3.1 shows the image of 24×24 correlation matrix of the wine spectrum data in grey scale. Here white color represents 1, black color represents 0, and grey color represents the correlation coefficient between 0 and 1. The diagonal elements are all white because their self-correlation coefficients must be 1. In this data, the correlation coefficients are found to range from at least .9 to 1. There is a certain pattern in the image, and it repeats every four elements, which is actually the number of wine types used in the experiment. Therefore, it suggests that there be a certain number of groups in the data.

4. Clustering Spectra Using the Cross-Correlation and the Pseudo-F

As an unsupervised learning method (Duda *et al.*, 2001; Hasite *et al.*, 2001; Johnson and Wichern, 2007; Krzanowski, 2000; Rencher, 2002), clustering can recognize the pattern by grouping the spectra which are close in terms of their synchronization. Among many possibilities, we use AHC because it is intuitive and easy to look at the latent structural relationship of all the spectra. One of the big advantages of AHC is that there is no need to assume the number of clusters which is

usually unknown in the first place.

To cluster the n wine spectra, a series of merging takes place from n clusters to a single cluster. At each stage of the hierarchical clustering the two closest or most similar clusters are linked together. The different ways of defining similarity or dissimilarity between clusters would develop different clusters. We have used the wavelength cross-correlation as a distance metric to catch the simultaneous movement of two spectra. Often, different linkages can also lead to different clusters, their performances turns out to be the same in the wine data since clusters are compact and well separated.

Among various criterion functions to evaluate the performances of the clustering methods (Choi and Jun, 2007; Milligan and Cooper, 1985), we use the pseudo-F which is intuitive to understand and has often big gaps or local maxima. Let us first define the within-variance and the between-variance among clusters. Suppose that for a spectrum $s_i \in R^p$, $i = 1, \ldots, n$, let the data be a set of $D = \{s_1, s_2, \ldots, s_n\}$ and cluster them into c disjoint clusters, $D_1, D_2, \ldots, D_c$. Let n_i be the size of D_i. Let $m_i = \sum_{s \in D_i} s/n_i$ be the mean of the cluster D_i, and $m = \sum_{s \in D} s/n$ be the grand mean. Let the variance V_i of the cluster D_i be

$$V_i = \sum_{s \in D_i} (s - m_i)(s - m_i)^T. \tag{4.1}$$

Let the total variance V_T be the sum of V_W and V_B, where

$$V_W = \sum_{i=1}^{c} V_i \quad \text{and} \quad V_B = \sum_{i=1}^{c} n_i (m_i - m)(m_i - m)^T. \tag{4.2}$$

Then $\mathrm{tr}(V_W)$ is called the within-variance, and $\mathrm{tr}(V_B)$ is called the between-variance. Since the total variance is fixed for a given data set, the within-variance and the between-variance are complementary. Often, the within-variance is called the sum of squared error(SSE) and used as a popular criterion to estimate the number of optimum clusters (Duda *et al.*, 2001; Choi and Jun, 2007). However, SSE considers only the within-variance and often decreases monotonically without a big gap or local minima, which leads to a hard decision.

Avoiding any parametric assumptions while keeping the fine property of two variances, we use the pseudo-F, an analog of the univariate analysis of variance, defined as follows (Rencher, 2002):

$$\text{pseudo-}F = \frac{\mathrm{tr}(V_B)/(c-1)}{\mathrm{tr}(V_W)/(n-c)}. \tag{4.3}$$

Note that it is a ratio of the between-variance to the within-variance, each divided by its corresponding degrees of freedom. When the between-variance becomes big, the clusters are well separated, and the within-variance becomes small, the clusters are compact. So the pseudo-F becomes bigger when the clusters are more compact and better separated. It approximately follows an F-distribution and often has local maxima, which is its very unique property. We estimate the number of clusters at its first local maximum.

Figure 4.1 illustrates the dendrogram of the hierarchical clustering with the wine data. If the tree are cut at a certain height, the data are separated into a number of clusters. From the structure, we can see that 2, 3 or 4 are most likely number of clusters. $\mathrm{tr}(V_B)$, $\mathrm{tr}(V_W)$ or SSE, and pseudo-F have been evaluated from the wine data. Luckily, both SSE and pseudo-F have local extremes, which is not the usual case. However as explained earlier we suggest to look at the pseudo-F because it considers both the between-variance and the within-variance, and mostly it has local maxima.

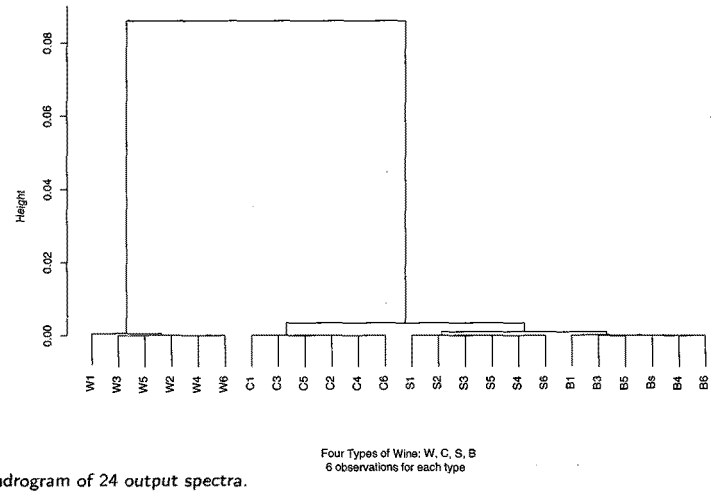

Figure 4.1. Dendrogram of 24 output spectra.

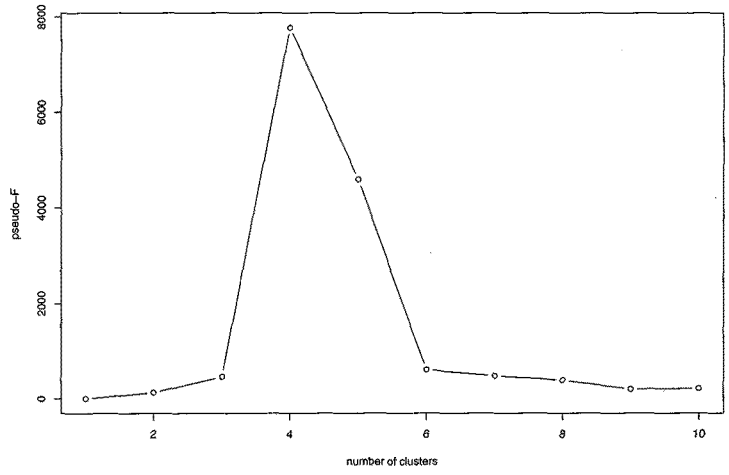

Figure 4.2. Pseudo-F for a number of cluster.

Figure 4.2 illustrates a plot of the pseudo-F for a serial number of clusters. Clearly, there is a maximum at 4, which is the very number of wine types used in the experiment. Above all, the four clusters match the four original types of wine. This empirical study announces that the well-made miniature spectrometer of filter-array provides reliable measurement of liquid materials.

5. Classification Based on the Cross-Correlation with the Mean

Once the materials have been clustered into a certain number of clusters, we are interested in whether the spectrometer can classify a new sample into one of the current clusters. For this supervised learning (Duda *et al.*, 2001), we consider a combination of two intuitive classification methods: the Nearest-Neighbor Rule(NNR) and Fisher Linear Discriminant Analysis(FLDA). While the NNR

Table 5.1. Between-variance $\mathrm{tr}(V_B)$, within-variance $\mathrm{tr}(V_W)$ which is SSE and pseudo-F for a given number of clusters obtained from AHC

Number of clusters	$\mathrm{tr}(V_B)$	$\mathrm{tr}(V_W)$	Pseudo-F
2	62732.24	10858.92255	127.0945
3	71923.79	1667.37509	452.9274
4	73528.07	63.09726	7768.7538
5	2015014.02	2084.46971	4591.7274
6	3784644.15	22419.92843	607.7057
7	3784649.26	22414.82421	478.3965
8	3784650.35	22413.73008	385.9522
9	5654073.17	53318.65901	198.8307
10	7587089.59	55541.71255	212.4915

assigns a test point to a group associated with the nearest point to it, FLDA assigns a test point to a group whose standardized mean is closer to it. Since a spectrometer is small, a simple algorithm is preferred. The NNR requires n comparison, and FLDA requires inversion of $p \times p$ covariance matrix where p is as big as the number of filters. In this work, we suggest that a new sample be assigned to a cluster whose mean has the highest cross-correlation.

Using the leave-one-out method, the 24 wine data are classified to one of the four wine types where it has the highest correlation with the mean of each type. Table 5.1 shows the correlation coefficients between observations and means of 4 clusters have been evaluated, and 100% of the data have been assigned to its own type. This good result implies that the cross-correlation is a fine metric to measure the relationship among spectra. Also, the spectrometer of low-cost and filter-array produces very reliable measurement even though its filter is not exactly a delta function.

6. Conclusion

A fine miniature spectrometer of low-cost and filter-array can detect liquid materials even though filters are not an exact delta function in a narrow range of response spectra. The distorted response spectra can be controlled by DSP because the output spectra have the manageable random variation even though spectra are distorted and augmented by the responses from other adjacent filters. Without recovering the original target spectrum, both unsupervised and supervised learning methods based on the wavelength cross-correlation have showed high performance in detecting the material types of the experimental data. This in turn implies that the spectrometer provides very similar distorted curves and so its measurement is reliable in terms of its repeatability. In order to use the cross-correlation as a metric for both unsupervised and supervised learning methods any dimension reduction or feature extraction methods have not been used only because the cross-correlation would be lost. If there is a necessity to reduce the dimension of the data, it is recommended to choose a series of consecutive filters instead of using PCA, ICA, RRR, or CCC as long as the correlation is considered as a metric to measure the distances between spectra.

The actual situation can make the conditions change, so that the DSP methods should be altered too. First, if liquid materials are poured into various containers of different materials, thickness or colors, the spectra could shift or delay across wavelengths. For those cases, the wavelet shift as an analogy of time shift should be introduced to the correlation. Secondly, the correlation coefficients used in this paper is often used for parametric studies when the distribution of data is assumed. If there

Table 5.2. Cross-correlation coefficients between an observation and the means of clusters. Each observation is assigned to a cluster whose mean has the highest correlation with it.

Obs	Original Cluster	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Assigned Cluster
1	1	0.9998216	0.9151263	0.9980889	0.9989154	1
2	2	0.9068464	0.9994405	0.8901063	0.9206034	2
3	3	0.9979018	0.8975372	0.9999163	0.9943889	3
4	4	0.9989331	0.9287489	0.9949036	0.9999427	4
5	1	0.9999558	0.9154419	0.9982293	0.9988450	1
6	2	0.9165701	0.9999625	0.9003364	0.9298916	2
7	3	0.9983196	0.8993117	0.9999911	0.9950370	3
8	4	0.9989958	0.9278646	0.9950127	0.9999896	4
9	1	0.9999517	0.9140040	0.9983476	0.9988785	1
10	2	0.9148905	0.9999871	0.8985333	0.9282904	2
11	3	0.9980765	0.8979680	0.9999877	0.9946364	3
12	4	0.9989880	0.9285916	0.9949886	0.9999926	4
13	1	0.9999520	0.9158218	0.9981637	0.9988977	1
14	2	0.9192420	0.9998583	0.9031421	0.9324349	2
15	3	0.9984355	0.8998715	0.9999838	0.9952306	3
16	4	0.9989372	0.9281958	0.9948908	0.9999888	4
17	1	0.9999444	0.9146531	0.9982193	0.9989663	1
18	2	0.9142444	0.9999614	0.8978503	0.9276916	2
19	3	0.9981283	0.8980259	0.9999849	0.9947162	3
20	4	0.9989093	0.9291263	0.9948338	0.9999901	4
21	1	0.9999511	0.9156939	0.9981371	0.9989158	1
22	2	0.9181948	0.9999058	0.9020335	0.9314520	2
23	3	0.9985161	0.9002765	0.9999693	0.9953614	3
24	4	0.9988707	0.9287655	0.9947574	0.9999851	4

are at least one outlier exists in the data, the coefficient breaks down. So, if it is necessary to avoid the parametric assumptions or break-down situation, the nonparametric correlation coefficients can be substituted as new metrics. Thirdly, if the correlation coefficients do not work satisfactorily as a metric, distances like Euclidean can be used along with typical dimension reduction methods. Alternatively, since the output spectra look like a function or density, the nonparametric methods to test the distribution can be used as a new metric, and this can provides a nice asymptotic distribution theory.

References

Box, G. E. P., Jenkins, G. M. and Reinsel, G. (1994). *Time Series Analysis: Forecasting and Control*, Wiley Series in Probability and Statistics, San Francisco.

Chang, C. C. and Lee, H. N. (2008). On the estimation of target spectrum for filter-array based spectrometers, *Optical Express*, **16**, 1056–1061.

Choi, K. and Jun, C. (2007). A systematic approach to the Kansei factors of tactile sense regarding the surface roughness, *Applied Ergonomics*, **38**, 53–63.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001). *Pattern Classification*, Wiley & Sons, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer-Verlag, New York.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*, Prentice Hall, New York.

Krzanowski, W. J. (2000). *Principles of Multivariate Analysis: A User's Perspective*, Oxford University Press, Oxford.

Milligan, G. W. and Cooper, M. C.(1985). An examination of procedures for determining the number of clusters in a data set, *Psychometrika*, **50**, 159–179.

Morawski, R. Z. (2006). Spectrophotometric applications of digital signal processing, *Measurement Science Technology*, **17**, 117–144.

Peebles, P. Z. (2000). *Probability, Random Variables and Random Signal Principles*, McGraw-Hill, New York.

Rencher, A. C. (2002). *Methods of Multivariate Analysis*, Wiley Series in Probability and Statistics, New York.