APPLICATION OF NON-NEGATIVE MATRIX FACTORIZATION TO FLUORESCENCE SPECTROSCOPY

Cyril Gobinet, Eric Perrin, and Régis Huez

Laboratoire d'Automatique et de Microélectronique, Université de Reims Champagne-Ardenne
Campus du Moulin de la Housse, B.P. 1039, 51687 REIMS Cedex 2, FRANCE
phone: +33.(0)3.26.91.82.21, fax: +33.(0)3.26.91.31.06, email: cyril.gobinet@univ-reims.fr

ABSTRACT

This article deals with application of signal processing and chemometric techniques to fluorescence spectroscopy. Recorded spectra of pure components in this field are characterized by very large peaks and come from a mixture of pure elements. It may be quite difficult to reconstruct the pure components spectra because of their mutually statistically dependence. We have decided to analyse existing techniques to resolve this problem.

1. INTRODUCTION

An important problem in chemistry and environmental sciences is the so called inverse problem. During a chemical reaction between different chemical species, a difficult problem is to observe the spectra and the concentrations profiles of intermediate species that can be formed during the reaction. The only available informations are spectra and concentrations of reactants and resultant products.

The first method that comes to mind is to record the spectra of pure reactants and products, and to substract them from the spectrum of the mixtures by the use of a regression method. But pure chemical species are difficult to obtain because of impurities that cannot be totally excluded. Furthermore, pure chemical species are not always available, being mixed with other components.

It is thus necessary to develop techniques that are not depending on knowledge of pure chemical species spectra.

Since 1971 with Lawton and Sylvestre [1], pure component spectra reconstruction is an important field of research. The aim is to estimate spectra of initial components from an observed set of additive mixtures. Their method was limited to the case of mixtures of two components. But their geometrical approach was simple and based only on the use of postitive constraints and normalisation of spectra. But good estimation results encouraged people to continue this initial work.

Among those persons, Ohta [2] extended this method to a three components system by the use of a MonteCarlo technique. But it was still not able to process a multidimensionnal problem.

Sasaki, Kawata and Minami [3] were the first to propose a theoretical extension to this multicomponent problem. Their work is based on the use of maximization of entropy with positivity constraints. In practice, only two or three components problems lead to a good estimation of pure spectra.

The main drawback of those methods is their non-single solution philosophy. They return a band of possible solutions. Depending on the variety of the mixtures and the level of noise, band may be very large. The user is unable to predict the spectra of pure species.

Several methods have been developed, for instance by Malinowsky [4] or Gemperline [5]. However they are not the widely used in spectroscopy from where the data used in this article come.

Thus, in section 2, data, that need to be processed, will be introduced, and their properties extracted. From the last, potentially useful and widely used algorithms will be discribed in section 3, and their drawbacks analysed. On this base, Non-negative Matrix Factorization (NMF) will be shown in section 4 as being the most effective technique to process data used in this article, and results will be presented. Section 5 will conclude this article.

2. FLUORESCENCE SPECTROSCOPIC DATA

2.1 Experimental considerations

Laser scanning microspectrofluorometry is used to collect the fluorescence signals. Samples of *durum wheat* grains were chosen from a serie of *triculum durum* used for evaluating the milling efficiency at INRA Montpellier (France). Transverse sections (60 µm) of the wheat grain were obtained by soaking the grains 4 hours in distilled water and then embedded in ice in order to be sectioned with the freezing microtome, which is an instrument to cut thin slices of a frozen sample. The microspectrofluorometer is equipped with a laser excitation at 365 nm and fluorescence signal emitted by the sample is collected on a CCD detector in the spectral interval 350 to 670 nm. In order to obtain an image, the laser scans an area of a several μm^2 in a point by point mode at a spatial resolution of about 1 µm. A 20×20 measurement matrix can thus be obtained ; each measurement consists of a 350 to 670 nm fluorescence spectrum. The original spectra are shown in figure 1. For clarity reasons, only 20 spectra randomly chosen among 400 are presented in the figure 1.

The figure 2 represents the same spectra, but they are normalized. It is easy to point out that the low level spectra have a high level of noise. That is a difficulty to extract the sources without uncertainty.

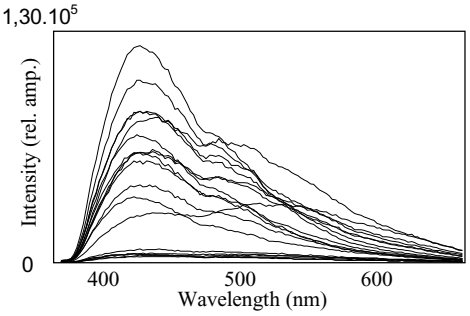

Figure 1 : Original spectra.

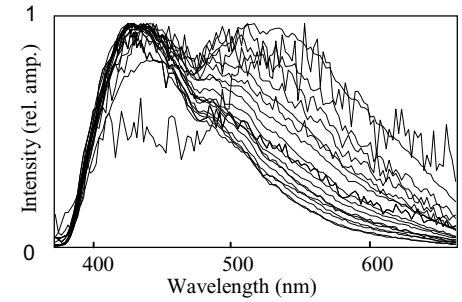

Figure 2 : Normalized spectra.

2.2 Data model description

According to physical laws governing fluorescence spectroscopy, the spectrum of a chemical species results from the weighted sum of the spectra of the pure components. The weights are the concentration of each pure species in the mixtures. Moreover, no transmission delay is assumed. The data model is assumed to be instantaneous and linear :

$$X = A\,S \tag{1}$$

where X is a $(N \times L)$ observed data matrix, S is a $(M \times L)$ unknown sources matrix, A is a $(N \times M)$ unknown mixing matrix corresponding to the concentration of each source in mixtures, N is the number of observed mixture spectra, M is the number of unknown sources and L is the number of points in each spectrum.

A physical property of spectra is available and will be very useful in the following to identify S and A : the positivity of entries of matrices S and A. Thanks to those constraints, the space of the solutions will be reduced. Different approaches have been developed to deal with this factorization problem. The aim is to find the matrices S and A that best fit the model described by equation (1).

3. POTENTIALLY USEFUL ALGORITHMS

3.1 MDF algorithm

A biophysics group processed those fluorescence data with the help of their own algorithm : the Maximum Distance Factor (MDF) algorithm [6]. It is based on a deflation approach. Original spectra with maximal distance between them are considered as pure component spectra. Of course they satisfy the non-negativity restriction for intensity values, but do not for concentrations. An expanded procedure with iterative correction of the obtained spectra has been developed, leading to positive concentrations. Constraints are then respected.

Good results have been obtained by application of MDF to fluorescence data, as can be seen in figure 3. Specialists can easily recognize the ferulic acid spectrum (n°1), the free ferulic acid spectrum (n°2) and the p-coumaric acid spectrum (n°3) which are represented in this figure.

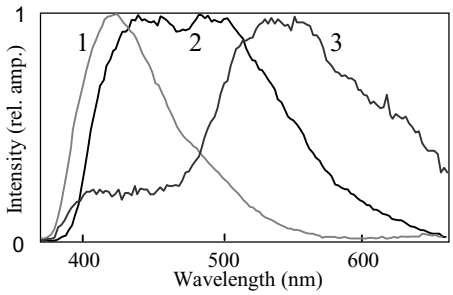

Figure 3 : Pure spectra estimated by MDF.

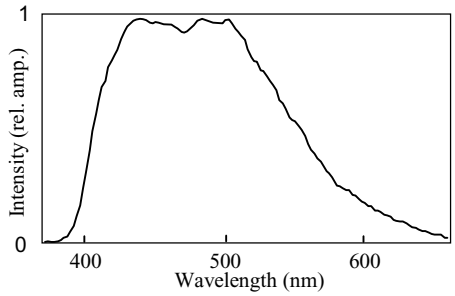

Figure 4 : Normalized fluorescence spectrum of free ferulic acid.

The effectiveness of this algorithm is observable thanks to the pure free ferulic acid spectrum that was measured independently of this experiment and that is available in figure 4. Comparison of the pure free ferulic acid spectrum and the estimated free ferulic acid spectrum confirms the good pure spectra extraction.

Nevertheless, two drawbacks are noticeable. First, practically pure spectra of chemical species must be present in the data matrix X in order to provide good results. Second, the mathematical framework is not as straightforward as it could seem and involves heavy computational time.

3.2 ICA algorithms

The classical techniques of Independent Component Analysis (ICA) [7, 8] have already been successfully applied to Raman spectroscopy [9]. But on the wheat grain spectra, the fundamental assumption (mutual statistical independence of sources) is not fulfilled. We know that pure component spectra are quite similar, involving dependence structure between them. Application of standard ICA algorithm is useless. Wrong results will be sure obtained.

However, a solution may appear with the transposition of the problem. If spectral sources are dependent, why not assume the independence of concentrations? This interesting idea is unfortunately not realizable. The chemical species in a wheat grain are not randomly distributed. A wheat grain has a physical structure with different concentration areas. In fact,

knowing the concentration of one species allows to predict the concentration of the other ones. A dependence structure is still apparent.

Of course non-negative constraints [10] should be used in order to reduce the space of the solutions, but it would lead to unrealistic solutions, as the underlying sources need to be still statistically independent.

An interesting algorithm is the one proposed by Cichocki and Georgiev [10]. It assumes that the sources are dependent except for at least one frequency band. A standard ICA algorithm can be applied in this band of frequencies. But the difficulty of this technique is the localization of the frequency domain in which the sources are mutually independent. Such an assumption is not conceivable for spectral data processed in this paper.

3.3 PMF algorithm

The Positive Matrix Factorization developed by Paatero [11] is another alternative to resolve the problem. It is based on the minimization of the Frobenius norm of the modelling error with constraints of positivity of elements of matrices S and A. It has been shown that this optimization problem is equivalent to a weighted least squares problem. This algorithm has been very successful in environmental sciences and numerous applications are attributable to it.

Nevertheless, three drawbacks are noticed:

- The standard deviation of each entry of X needs to be known or to be quite well estimated by the user. In our case, this information is not available. An estimation of the standard deviation requires several experiments and the spectroscopic techniques are based to a single and fast measure.
- Even if the global maximum is reached, a rotation of the solution is still left since

  $AS = ATT^{-1}S$ for every non-singular matrix T.

  The user must specify this rotation to find the relevant solution.
- The computational cost is very expensive.

All that was said before leads us to use a technique based on positivity of matrices S and A, but that avoids the drawbacks of methods mentionned above. Our choice was made on a new method developed by Lee and Seung and based on the Non-negative Matrix Factorization (NMF).

4. NON-NEGATIVE MATRIX FACTORIZATION

4.1 Definition

Assuming there is no noise, the goal is to find non-negative matrix factors A and S that fulfilled the data model (1). The NMF methods only assume that the model spectra are non-negative and allow only additive combinations in the matrix A of the concentrations [12].

The constraint of non-subtractive combinations is of a great importance, because it corresponds more to the classical idea of the mixing of physical components. Each observation is only a positive weighted sum of the model spectra.

We have to define a cost function that quantifies the quality of the approximation. The simplest way is to measure the Euclidian distance between the two matrices X and $A.S$ [13]. This distance is defined by:

$$D = \|X - AS\|^2 . \qquad (2)$$

Lee and Seung have proposed some "multiplicative update rules" for A and S with a good compromise between speed and easiness of implementation for solving this minimization problem. Those rules are transcribed in equation (3) :

$$S_{ij} \leftarrow S_{ij} \cdot \frac{\left(A^T X\right)_{ij}}{\left(A^T AS\right)_{ij}} \quad and \quad A_{ki} \leftarrow A_{ki} \cdot \frac{\left(XS^T\right)_{ki}}{\left(ASS^T\right)_{ki}} . \qquad (3)$$

The Euclidian distance is not the only cost function; an other one based on the Kullback-Leibler divergence may be used. Nevertheless, simulations give similar results, and the choice of the Euclidian distance was arbitrary.

4.2 Application of NMF to fluorescence spectroscopy

Tests have been run on data described in section 2 with a sources number which is variable. Wrong results were obtained with two or four sources. A model with three sources gave a perfect restitution of the spectrum of free ferulic acid.

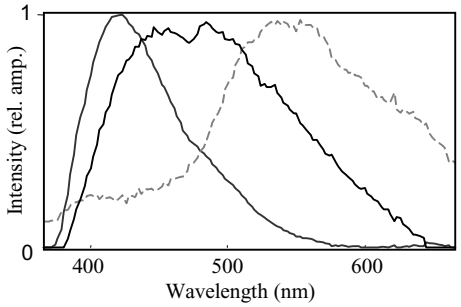

Figure 5 : Pure spectra estimated by NMF.

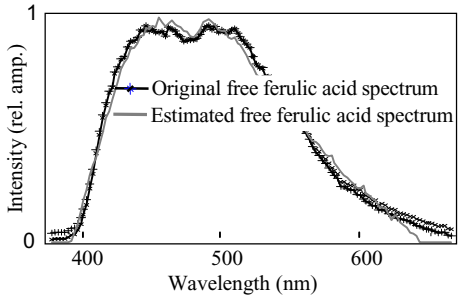

Figure 6 : Comparison of original and estimated spectra.

Figure 5 shows estimated pure spectra obtained by application of NMF with a 3 sources model. Results are similar to those obtained with the MDF algorithm (figure 3).
Figure 6 compares normalized a priori free ferulic acid spectrum with estimated free ferulic spectrum.

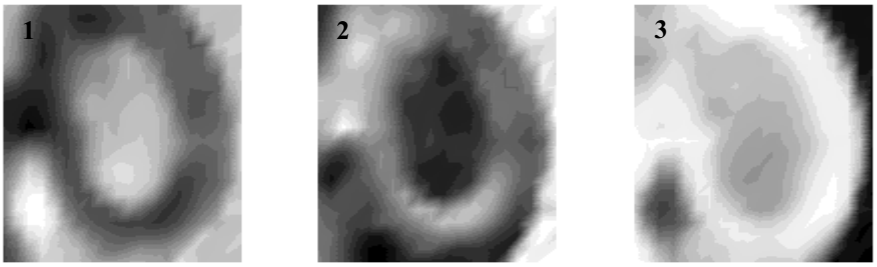

Figure 7 : Distribution and concentration of the chemical species in a wheat grain section (1) bound ferulic acid, (2) free ferulic acid, (3) p-coumaric acid.

After the determination of the three chemical species, their distribution and concentration can be visualized in the wheat sample using chemical maps. Each image corresponds to a column of the estimated mixing matrix. Figure 7 shows the spatial distribution of the different species. The concentration scale decreases from black to white. It can be noticed that the bound ferulic acid is concentrated at the periphery of the wheat grain, while its free form is mainly at the middle of the grain. As one can see on figure 7, the correlation between the concentrations is the reason why the ICA techniques don't give good results on the transposed data. The NMF method does not take into account the independence of the signals in the search for solutions, but is based on the positivity of estimated spectra intensity and estimated components concentrations, and thus seems to be more suited for the study of this type of signals.

5. CONCLUSION

In chemical terms, the results of the NMF methods, applied to the characterization and fluorescence chemical mapping of wheat grain sections, show that the first species characterizes the aleurone layer. It can be used as a meaningful indicator of the non-endosperm tissues of the grain in order to characterize and estimate aleurone contamination in different mill-streams.

The NMF method has shown its effectiveness to deal with problems that can not be easily solved by the application of the classical methods of BSS. This opens the way to an interesting and innovative way of research which relates to the introduction of positivity constraints into the procedure of sources separation.

REFERENCES

[1] W. H. Lawton and E. A. Sylvestre, "Self modeling curve resolution", *Technometrics*, vol. 13, pp. 617–633, 1971.

[2] N. Ohta, "Estimating absorption bands of component dyes by means of principal component analysis", *Analytical Chemistry*, vol. 45, pp. 553–557, 1973.

[3] K. Sasaki, S. Kawata and S. Minami, "Constrained nonlinear method for estimated component spectra from multicomponent mixtures", *Applied Optics*, vol. 22, pp. 3599–3603, 1983.

[4] E. R. Malinowski, "Obtaining the key set of typical vectors by factor analysis and subsequent isolation of component spectra", *Analytica Chimica Acta*, vol. 134, pp. 129–137, 1982.

[5] P. J. Gemperline, "Target transformation factor analysis with linear inequality constraints applied to spectroscopic-chromatographic data", *Analytical Chemistry*, vol. 58, pp. 2656–2663, 1986.

[6] S Charanov, A. Saadi, A. Kokota and M. Manfait, "Estimating spectral models for multidimensional data mapping", in *Proc. SPIE BIOS 1999*, San Jose, California, 1999, vol. 3605, pp 317–324.

[7] J.-F. Cardoso and A. Souloumiac, "Blind beamforming for non gaussian signals", *IEE Proceedings-F*, vol. 140, pp. 362–370, Dec. 1993.

[8] A. Belouchrani, K. Abed Meraim, J.-F. Cardoso and E. Moulines, "A blind source separation technique using the second order statistics", *IEEE Transaction on Signal Processing*, vol. 45, pp. 434–444, 1997.

[9] R. Huez, E. Perrin, G. D. Sockalingum, and M. Manfait, "Blind source separation, application to microorganism Raman spectra", *Proc. EUSIPCO 2002*, Toulouse, France, Sept. 2002, vol. 3, pp. 415–418.

[10] A. Cichocki and P. Georgiev, "Blind source separation with matrix constraints", *IEICE Trans. Fundamentals*, vol. E86-A, pp. 522–531, 2003.

[11] P. Paatero, "Least squares formulation of robust non-negative factor analysis", *Chemometrics and Intelligent Laboratory Systems*, vol. 37, pp. 15–35, 1997.

[12] D. D. Lee and H. S. Seung, "Learning the part of objects by non-negative matrix factorization", *Nature*, vol. 401, pp. 788–791, 1999.

[13] D. D. Lee and H. S. Seung, "Algorithms for non negative matrix factorization", *NIPS*, vol. 13, pp. 556–562, 2000.