ABOUT IMPORTANCE OF POSITIVITY CONSTRAINT FOR SOURCE SEPARATION IN FLUORESCENCE SPECTROSCOPY

Cyril Gobinet, Abdelkamel Elhafid, Valeriu Vrabie, Régis Huez, and Danielle Nuzillard

Centre de Recherche en Sciences et Technologies de l'Information et de la Communication
Université de Reims Champagne-Ardenne
Campus du Moulin de la Housse, B.P. 1039, 51687 REIMS Cedex 2, FRANCE
phone: +33(0)326918221, fax: +33(0)326913106
email: cyril.gobinet,abdelkamel.elhafid,valeriu.vrabie,regis.huez,danielle.nuzillard@univ-reims.fr

ABSTRACT

The importance of positivity constraint in source separation techniques for spectroscopic applications is presented in this paper.
Microspectrofluorometry measures fluorescence signals emitted by the analyzed tissues. The information associated to each pure chemical species must be estimated in order to characterize these tissues. Source separation techniques are well suited to this task. However, pure species spectra and concentrations non-negativity must be considered to obtain a realistic solution.
Applications of fluorescence spectroscopy on wheat and barley grains are analyzed. Each one has specific properties suggesting the use of two conceptually different algorithms: Non-negative Matrix Factorization (NMF) and Second Order Blind Identification followed by a positive procedure ("positive SOBI"). We show that no complementary experiments are needed to identify chemical species of the analyzed tissues.

1. INTRODUCTION

Microspectrofluorometry highlights the vibrational states of molecules by measuring at different wavelengths the intensity of light interacting with the analyzed tissues. The measure in one point, indexed by the wavelength, provides a spectrum. Chemical species are mixed with different proportions in each measure point. Dataset can be obtained when several measures are realized at different points. Recorded spectra contain non interpretable information about pure compounds and their concentrations. A recurrent problem in biophysics is obviously to separate individual information from collected spectra. On the one hand pure species spectra need to be estimated in order to identify these species, and on the other hand concentration profiles are requested to analyze species repartition in the sample. Whatever is the chosen algorithm to solve this source separation problem, positivity of pure species spectra and concentration profiles must be incorporated to ensure convergence to a physically meaningful solution. To illustrate this, two problems of source separation in fluorescence spectroscopy on wheat and barley grains are solved by different approaches.

Since 1970, researches have carried through development of several source separation approaches based on Principal Component Analysis (PCA). Chemistry and environmental sciences are the first fields which have initiated source separation researches by work of Lawton and Sylvestre [1]. An efficient separation is obtained for a two sources application. Extensions to more than two sources have been developed by Malinowski [2], and Windig *et al* [3]. As pure species spectra are positive and often overlapped, the decorrelation assumption is not justified. Due to the positivity, more realistic constraints and assumptions have been taken into account. Paatero [4] resolves a weighted least square formulation of a non-negative factor analysis problem. The work of Chew *et al* [5] is based on a transformation of the right matrix given by the singular value decomposition of the dataset. This transformation yields to the simplest spectrum associated to a characteristic spectral band. Positivity and intensity constraints are added to ensure convergence to a realistic solution. Nevertheless, these methods assume *a priori* knowledge as the variance of each recorded spectra for each wavelength, or characteristic spectral bands of unknown sources.

Since 1990, signal processing community investigates the blind source separation methods [6]. The main result is the development of Independent Component Analysis (ICA) [7], which relies on mutual statistical independence of underlying sources. Positivity constraint-based extensions for higher-order identification methods have been recently developed [8, 9]. Since most fluorescent molecules have very large and unstructured fluorescent bands [10] and spectra have positive intensities, the independence can't be assumed. It will be shown that second-order independence and lag-dependent correlation of concentration profiles are more realistic assumptions, leading to the use of Second Order Blind Identification (SOBI) [11] for the barley grain analysis. SOBI is a very popular algorithm in spectroscopic applications such as Nuclear Magnetic Resonance, Raman and Infrared Spectroscopy, which are described in [12, 13, 14]. As positivity is essential for our applications, a procedure to force the positivity of sources and mixing matrix is added, leading to a "positive SOBI" algorithm. For the wheat grain analysis, the second order independence is shown to be an unrealistic solution. In this case of positive mixtures of positive sources, a specific algorithm named Non-negative Matrix Factorization (NMF) [15, 16] can be employed. It was proved that this technique is very efficient [17], even for the barley grain analysis.

The purpose of this article is to show that quite similar problems of source separation can be solved by different approaches, and that positivity constraint is essential as soon as spectroscopic data need to be processed. In section 2 we describe two fluorescence spectroscopy applications on wheat and barley grains. Based on their characteristics, algorithms used for each application are briefly described in section 3. These algorithms are the NMF for the wheat grain application and "positive SOBI" for the barley grain study. Section 4 presents results of their applications. Finally, we conclude in section 5.

2. TWO DIFFERENT PROBLEMS OF FLUORESCENCE SPECTROSCOPY

In order to illustrate the importance of positivity constraint, two different applications of fluorescence spectroscopy are studied.

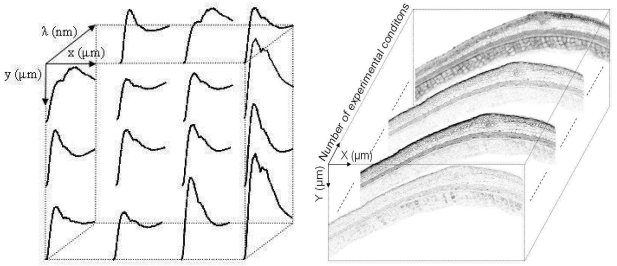
2.1 Wheat grain

Auto-fluorescence emission spectra from a transversal section of wheat grain were recorded using a confocal laser microspectrofluorometer equipped with an optical microscope. The excitation laser at 365 nm scans an XY area of several m^2 in a point by point mode. Each spectrum was recorded in 128 wavelengths in the spectral interval from 350 to 670 nm. A $20 \times 20 \times 128$ data cube is obtained. For clarity reasons, only few spectra are represented in figure 1(a). Recorded spectra contain mixed information of phenolics, which are the most auto-fluorescent materials in a wheat grain. The aim is to separate pure molecular species information in order

to characterize them and to analyze their spatial distributions in a wheat grain. As the spectral resolution is good, a precise estimation of pure species spectra is hoped. In the opposite way, as only a 20×20 pixels image is used, the poor spatial resolution prevents the visualization of fine underlying biological structures. For biophysicists, a species will be found to be related with the aleurone layer. Aleurone contamination in different millstreams will thus be quantified thanks to this indicator species.

2.2 Barley grain

Confocal laser microscpectrofluorometry enables also to visualize autofluorescence of vegetal walls. A set of four excitation laser at 364, 488, 543 and 633 nm and a set of 9 emission filters define 19 acquisitions conditions leading to the recording of 19 spectral images of a barley grain [18]. As these acquisitions are made in long or band pass filters, recorded signals are indexed by the number of experimental conditions. Nonetheless, to simplify the notation and the comprehension, we denote them "spectra". Each image is composed by 512×512 pixels, each pixel corresponding to a physical point of the barley grain. A $512 \times 512 \times 19$ data cube is obtained. Few spectral images are shown in figure 1(b). Only external tissues are visible by fluorescence with these experimental conditions. The aim is to identify the different *in situ* tissues thanks to the identification of pure species and the estimation of their concentration profiles. Thanks to the good spatial resolution of data, biological structures will be isolated by estimation of concentration profiles of pure species. Nevertheless, as long or band pass filters are used, the spectral resolution is low. Pure species spectra won't be estimated with fineness. For biophysicists, a long term goal is to follow the evolution of tissues during transformation process such as grinding, flour or paste.



(a) Wheat grain (b) Barley grain

Figure 1: Data cubes

2.3 Datasets characteristics

In order to choose the well adapted algorithms, applications characteristics are studied. The first and obvious feature is the instantaneousness data recording because scattered light is collected by CCD detectors. Physical laws governing fluorescence spectroscopy mechanisms are well known to be linear. Recorded spectra thus result from weighted sum of spectra of pure species present in the analyzed tissues. This instantaneous and linear model is:

$$X = AS \tag{1}$$

where X is the data matrix obtained by concatenating consecutive lines of the recorded data cube, $X \in \mathbb{R}^{400\times128}$ for the first application and $X \in \mathbb{R}^{47021\times19}$ for the second one (because only 47021 from 512×512 spectra are not vanished). S is the source matrix of dimensions $M_1 \times 128$ respectively $M_2 \times 19$. A is the mixing matrix (or concentration coefficients) of dimensions $400 \times M_1$ respectively $47021 \times M_2$.

Positive recorded data X need to be expressed as a factorization of two positive matrices A and S. By positive matrix, we mean that each element of the matrix is positive. Estimated spectra and concentrations will thus be physically meaningful.

Since most fluorescent molecules have very large and unstructured fluorescent bands [10] and spectra have positive intensities, pure species fluorescent spectra don't fulfill mutual independence even at the second order. ICA methods are not appropriate to solve the equation (1), but they may solve the transpose problem $X^T = S^T A^T$. In blind source separation language, sources are now lines of A^T, i.e. pure species concentration profiles, and the mixing matrix is represented by S^T, i.e. each column of S^T being a pure species fluorescent spectrum. Mutual independence of pure species concentration profiles is now required. A biological analysis is thus needed to argue about independence of concentration profiles. As every biological element, a wheat or barley grain has a well organized structure, so pure compounds are not randomly distributed. A dependent structure between pure compounds concentration profiles exists, which implies that ICA algorithms based on higher order statistics are not adapted. However for second order independence and lag-dependent correlation of concentration distributions, SOBI-type algorithms [11] can be employed.

For the first application, a previous study [19] has shown that wheat grain pure species are diffused with overlapping concentration areas but also with zones where species don't exist simultaneously. This property implies that the matrix A (made up by concentration coefficients) is a sparse matrix. But, due to the positivity of the concentrations, the second order independence assumption is unrealistic, so SOBI-type algorithms are useless. In this case, the formulation of equation (1) is very similar to that given by Lee and Seung in [16], leading to the use of NMF algorithm.

The second application presents different characteristics. Barley spectroscopic data have been measured in a very small area of the external layer. Each point of measure has a much better spatial resolution than previously. The different tissues are very well identified and parted from the others. A preliminary study [18] shows that each tissue is almost composed by only one auto-fluorescent species. Second order independence and lag-dependent correlation of concentration profiles are thus almost fulfilled assumptions. Furthermore, the pure species are by definition primary compounds, so their associated spectra are linearly independent. As a consequence, SOBI-type algorithms can be used for this application. However, another *a priori* information must be exploited : the non-negativity of sources and mixing matrix.

Applications characteristics being known, a mathematical formulation of problems is possible.

Problem 1: Knowing a positive data matrix X, find a factorization by two positive matrices A and S such that $X \approx AS$ and that A is a sparse matrix.

Problem 2: Knowing a positive data matrix X^T, find two positive matrices S^T and A^T solving $X^T = S^T A^T$, such that columns of S^T are linearly independent and lines of A^T are second order independent.

Both problems are now well defined. In the following we present the suitable algorithms to resolve them.

3. WELL SUITED ALGORITHMS

3.1 Non-negative Matrix Factorization

Pure species spectra are mutually dependent, but also pure species concentration profiles, so ICA algorithms are useless. The formulation of problem 1 is similar to that given by Lee and Seung in [16] to describe the goal of NMF. Our choice thus turns towards this algorithm.

The required assumption is the positivity of factorization matrices. Based on it, a closely related to the Kullback-Leibler divergence objective function is used to derive a simple and very efficient optimization algorithm. Multiplicative update rules of $S = \{S_{ij}\}$ and $A = \{A_{ki}\}$ are mathematically expressed by

$$S_{ij} \leftarrow S_{ij} \frac{{}_k A_{ki} X_{kj} / (AS)_{kj}}{{}_l A_{li}} \tag{2}$$

$$A_{ki} \leftarrow A_{ki} \frac{\sum_j S_{ij} X_{kj} / (AS)_{kj}}{\sum_m S_{im}}. \quad (3)$$

This algorithm has been proved to converge to a local minimum of optimized cost function. One of its advantage is that it is free from any step size parameter.
As said, this method will be applied to wheat grain.

3.2 Second Order Blind Identification and positive procedure

Several ICA algorithms have been recently developed in order to insert the positivity constraint of sources [8, 9]. However, these algorithms exploit higher order cumulants. As mentioned, only second order independent and lag-dependent correlation of pure species concentration profiles can be assumed. Nevertheless, second order independence algorithms based on positivity constraint haven't been yet developed. That's why we use a two step method.
The first step consists to apply SOBI [11] to the data matrix X^T.
A positive procedure [13] is used in the second step. Negative values of the sources A^T are set to zero. A least mean square approximation estimates pure species spectra:

$$S^T = X^T A (A^T A)^{-1}. \quad (4)$$

The negative elements of the obtained mixing matrix S^T are set to zero. A new least mean square approximation is used in order to estimate pure species concentration profiles:

$$A^T = (SS^T)^{-1} S X^T. \quad (5)$$

This second step is repeated until convergence is reached.

A remark must be done about the legitimacy of the positive procedure. A pure species is the major compound of a tissue, but is also present in little concentrations in the other tissues. Second order independence of concentration profiles is not totally verified in the barley grain. Thus an estimation error is done by SOBI on A^T. Consequently, an error is also made on S^T. Even if these errors are not important, they can lead to negative matrix elements, which is in contradiction with physical constraints. A procedure to avoid negative values relaxes the second order independence assumption.

4. APPLICATIONS

4.1 NMF results on wheat grain

Each spectrum of the dataset X presented in section 2.1 has been normalized to unit area. This preprocessing step is useful to speed up convergence of NMF algorithm. The only adjusting parameter is the number of underlying sources. Data were then injected in the algorithm described by equations (2) and (3).

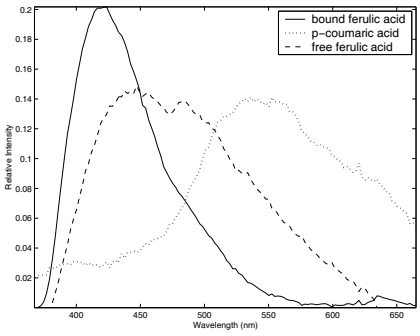


Figure 2: Pure species spectra estimated by NMF.

Principal Component Analysis suggests a three or four sources model. A preliminary study [19] and results analysis of biophysicists confirm good estimations obtained with $M_1 = 3$ pure species. As solutions are dependent from initialization of the algorithm, the mean solution over 100 trials has been computed. Resulting spectra are represented in figure 2. Expert analysis associated the solid line spectrum to bound ferulic acid, the dashed line to free ferulic acid and the dotted line to p-coumaric acid. Pure species concentration distributions can be viewed in figure 3 thanks to chemical maps. Each image corresponds to a column of the estimated mixing matrix A. The concentration scale decreases from black to white. It can be noticed that bound ferulic acid is concentrated at the periphery of the wheat grain, while its free form is mainly at the middle. P-coumaric acid is slightly present in wheat grain [17]. As mentioned, good spectral respectively poor spatial resolution are obtained for this application.

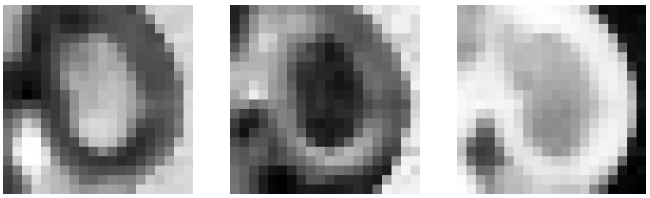


Figure 3: Pure species concentration profiles estimated by NMF. From left to right: bound ferulic, free ferulic, p-coumaric acids.

Thanks to these results, biophysicists concluded that the aleurone layer is characterized by bound ferulic acid.

4.2 "Positive SOBI" results on barley grain

Concentration profiles of barley grains have been shown to be second order independent and to possess a lag-dependent correlation. SOBI has been applied to dataset X^T presented in section 2.2. Whatever the number of sources, estimated sources and mixing coefficients exhibit some negative values. Positivity constraint as described in section 3.2 is required to force the solutions to positivity. But, as soon as iterative procedure of equations (4) and (5) is run, second order independence is no more completely fulfilled by new estimated sources. Consequently, a compromise must be done between second order independence and positivity of sources. Three parameters need to be fixed by the user. The number of sources is chosen equal to $M_2 = 4$ as suggested by a preliminary PCA step. The number of correlation matrices to be diagonalized is taken equal to 7, because results are slightly better. But this number doesn't have much influence on results. Positive procedure iterations are equal to 20 to ensure that each element of sources and mixing matrices are positive. Even if a tissue is constituted by a major species, other species may be present in small quantities as was said in section 2.3, so the positivity is preferred here to second order independence.

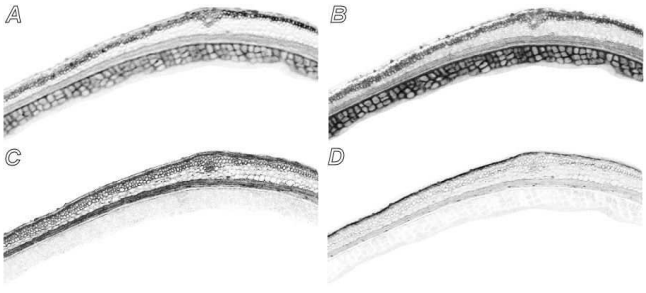


Figure 4: Spatial repartition of pure species estimated by "positive SOBI"

Estimated spatial repartitions of pure species are represented in figure 4. Different biological structures appear in these images. A specific layer is almost composed by a single pure fluorescent species, but it can be observed that images A and B are very related, just a scale intensity coefficient differs. A single chemical species

seems to have generated these two images. Identification of these species can be done thanks to the study of estimated pure species spectra, which are depicted in figure 5 (solid lines). As mentioned, these plots denoted "spectra" are indexed by the number of experimental conditions.

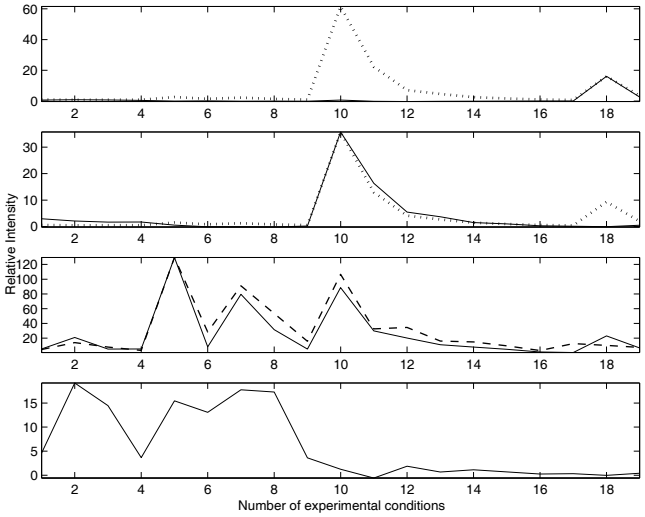


Figure 5: Pure compounds spectra estimated by "positive SOBI" (solid lines), ferulic acid (dotted lines) and lignin (dashed line)

It is well known that the most auto-fluorescent compounds are ferulic acid, lignin and cutin. Reference spectra of ferulic acid and lignin have been measured and are available in figure 5 (dotted respectively dashed lines). It is easily seen that a weighted sum of first and second spectra represented by solid lines in figure 5 is similar to the ferulic acid spectrum (dotted line). The observed correlation between image A and B of figure 4 is thus confirmed. The third spectrum represented by solid line is associated to the lignin spectrum (dashed line). Cutin is represented by the fourth spectrum of figure 5. Knowing pure species spectra and their concentration profiles, their biological localization can be discussed. Ferulic acid is present in majority in the aleurone layer and in the lemmae as can be seen on images A and B of figure 4. Cutin is concentrated in the top waxy layer as it is depicted on image D, while lignin is a major constituent of the pericarp and the bottom waxy layer of figure C. Each tissue is thus associated to a singular pure species. As suggested, good spatial respectively poor spectral resolution are obtained for this second application.

A remark must be done. The mathematical formulation of problem 2 suggests also the use of NMF. Results are identical to those exposed on figure 4 and 5. However, computational cost is more expensive. "Positive SOBI" is thus a well alternative to solve problems when computational time must not be too long, especially when data cube dimensions are large.

5. CONCLUSION

Two fluorescence spectroscopy applications have been described. Even if the final goal and experimental measures are quite similar, dimensions of datasets and physical respectively chemical characteristics of each application are different. Conceptually different source separation methods must thus be used. However, biophysicists need to have physically interpretable results. Positivity of concentration profiles and spectra of pure chemical species is thus required. Whatever is the chosen algorithm, this constraint force algorithms to converge to a realistic solution. A study of applications properties and the importance of positivity constraint convince us to exploit advantages proposed by NMF and "positive SOBI". No complementary experiments are needed to identify chemical species of the analyzed tissues.

REFERENCES

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

[2] 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.

[3] W. Windig, J. L. Lippert, M. J. Robbins, K. R. Kresinske and J. P. Twist, "Interactive self-modeling multivariate analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 9, pp. 7–30, 1990.

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

[5] W. Chew, E. Widjaja and M. Garland, "Band-Target Entropy Minimization (BTEM): an advanced method for recovering unknown pure component spectra. Application to the FTIR spectra of unstable organometallic mixtures," *Organometallics*, vol. 21, pp. 1982–1990, 2002.

[6] C. Jutten and J. Herault, "Blind separation of sources, part I: An adaptive algorithm based on neuromimietic architecture," *Signal Processing*, vol. 24, pp. 1–10, 1991.

[7] P. Comon, "Independent component analysis, a new concept?," *Signal Processing*, vol. 36, pp. 287–314, 1994.

[8] M. D. Plumbley and E. Oja, "A"nonnegative PCA" algorithm for independent component analysis," *IEEE Transactions on Neural Networks*, vol. 15, pp. 66–76, 2004.

[9] Z. Yuan and E. Oja, "A FastICA algorithm for non-negative independent component analysis," in *Proc. ICA 2004*, Granada, Spain, Sept. 22-24. 2004, pp.1–8.

[10] B. Valeur, *Fluorescence moléculaire*, De Boeck, Brussels, 2004.

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

[12] D. Nuzillard, S. Bourg and J.-M. Nuzillard, "Model-free analysis of mixture by NMR," *Journal of Magnetic Resonance*, vol. 133, pp. 358–363, 1998.

[13] D. Nuzillard and J.-M. Nuzillard, "Blind source separation applied to non-orthogonal signals," in *Proc. ICA 1999*, Aussois, France, Jan. 11-15. 1999, pp. 25–30.

[14] R. Huez, E. Perrin, G. D. Sockalingum and M. Manfait, "Blind source separation, application to microorganism Raman spectra," in *Proc. EUSIPCO 2002*, Toulouse, France, Sept. 3-6. 2002, pp. 415–418.

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

[16] D. D. Lee and H. S. Seung, "Algorithms for non-negative matrix factorization," *Advances in Neural Information Processing Systems*, vol. 13, pp. 556–562, 2001.

[17] C. Gobinet, E. Perrin and R. Huez, "Application of Non-negative Matrix Factorization to fluorescence spectroscopy," in *Proc. EUSIPCO 2004*, Vienna, Austria, Sept. 6-10. 2004.

[18] P. Courcoux, M.-F. Devaux and B. Bouchet, "Simultaneous decomposition of multivariate images using three-way data analysis: Application to the comparison of cereal grains by confocal laser scanning microscopy," *Chemometrics and Intelligent Laboratory Systems*, vol. 62, pp. 103–113, 2002.

[19] A. Saadi, I. Lempereur, S. Sharonov, J.-C. Autran and M. Manfait, "Spatial distribution of phenolic materials in durum wheat grain as probed by confocal fluorescence spectral imaging," *Journal of Cereal Science*, vol. 28, pp. 107–114, 1998.