INSTITUT DE
STATISTIQUE

UNIVERSITÉ CATHOLIQUE DE LOUVAIN

DISCUSSION
PAPER

0811

USE OF ICA ON HPCL-DAD DATA AND HIGH-ORDER STATISTICS TO AUTOMATICALLY ACHIEVE PEAK PICKING

DEBRUS, B., LEBRUN, P., CECCATO, A., CALIARO, G., GOVAERTS, B.,

OLSEN, B., ROZET, E., BOULANGER, B. and P. HUBERT

Use of ICA on HPLC-DAD data and high-order statistics to automatically achieve peak picking.

Benjamin Debrus[*,1a], *Pierre Lebrun*[*,a], *Attilio Ceccato*[b], *Gabriel Caliaro*[c], *Bernadette Govaerts*[d], *Bernard Olsen*[e],*Eric Rozet*[a], *Bruno Boulanger*[f], *Philippe Hubert*[a]

a) Service de Chimie Analytique, Université de Liège, Belgium
b) GlaxoSmithKline Biologicals, Belgium
c) Orailac Quality Solutions, Belgium
d) Institut de Statistiques, Université catholique de Louvain, Belgium
e) Analytical Chemistry, Eli Lilly Company, Indianapolis, United States
f) UCB Pharma SA, R&D Clinical Pharmacometrics, Belgium

SUMMARY

One of the major difficulties within the context of the fully automated development of chromatographic methods consists in the automated detection of the peaks coming from complex matrices such as multicomponent pharmaceutical formulations or stability studies of these formulations. The same problem can also occur with plant materials or biological matrices. This step is thus critical and time-consuming, especially when Designs of Experiments (DOE) are used to generate chromatograms. The use of DOE leads to maximize the changes of the analytical conditions in order to cleverly explore an experimental domain. Unfortunately, this generally provides very different and "uncontrolled" chromatograms which can be hardly interpretable, complicating picking and peak tracking. In this context, numerical signal processing methods such as Independent Components Analysis (ICA) was investigated to solve this problem. The ICA principle assumes that the observed signal is the resultant of several phenomena (known as sources) and that all these sources are statistically independent. ICA is able to estimate sources which most often seem judicious to represent the constitutive components of a chromatogram. In the present study, ICA was applied to HPLC-UV-DAD chromatograms and we showed that ICA allows differentiating noises and artifacts components from those of interest, by applying clustering methods based on high-order statistics computed on these components. Furthermore, on the basis of the described numerical strategy, it was also possible to rebuild a cleaned chromatogram easily legible. This represents a very significant advance towards our final objective, the fully automated development of liquid chromatography (LC) method.

---

[1] Corresponding author: b.debrus@ulg.ac.be, +32 4366 4324, Fax: +32 4366 4317
* These authors contributed equally to this work

1. INTRODUCTION

In many frameworks such as the automated development and optimization of analytical methods, the use of Design of Experiments (DOE) is of great interest to achieve the initial goals of a process. This requires several or many experiments that will allow the modeling of the studied responses. However, it also means that analytical conditions of a liquid chromatographic (LC) method must be highly modified in order to cover an experimental space, as large as possible. This leads to very different chromatograms and the peaks picking and tracking is often problematic, even if it is manually done by a confirmed analyst. The automation of this process is therefore one of the first problems to solve to achieve the fully automated development of LC methods.

In the literature, different approaches are already proposed. A mathematical treatment like deconvolution (one-dimensional or less frequently multidimensional) is a widely used technique to accurately estimate the overlapping of peaks. It only needs the number of peaks and some information about the peak shapes as basic input parameters in the computer assisted peak deconvolution procedure.[1-3] However, the knowledge of overlapped peak shapes is sometimes difficult to determine, and poor results can sometimes be obtained.[4-6]

Methods such as alternating least-square multivariate curve resolution (ALS MCR), mutual automated peak matching (MAP) or factor analysis (FA) have been recently developed and seem to be very efficient in the multidimensional numerical separation domain applied on HPLC-DAD data. However, examples with real data demonstrate that some work is still mandatory to achieve a perfect automated detection and matching of peaks[7-9].

Independent Component Analysis (ICA) is an interesting alternative that may be used to achieve these identifications[10-12]. ICA is used in more and more domains due to its ability to separate successfully many types of signals. In this context, ICA is a Blind Source Separation method (BSS). The term *blind* indicates that both the source signals, and the way the signals are mixed, are unknown. This processing method has been successfully applied to the analysis of mixed sounds, satellites signals, to biomedical signal processing problems such as electroencephalographic (EEG) data[13-15] or functional magnetic resonance imaging (fMRI) data[16], etc. Recent papers have shown that ICA can be applied in the field of chromatography for metabolites peaks detection[17] and to resolve the overlapping gas chromatographic-mass spectrometric (GC-MS) signals[18,19] or 3D-fluorescence spectroscopy[20]. The statistical method finds the independent components (sources) by maximizing the statistical independence of the estimated components. Non-Gaussianity, motivated by the central limit theorem, is one method for measuring the independence of the components, and can be measured, for instance, by approximations of negentropy. Mutual information is another popular criterion for measuring statistical independence of signals. Typical algorithms for ICA use centering, whitening and dimensionality reduction as preprocessing steps in order to simplify and reduce the complexity of the problem for the algorithm. Algorithms for ICA include JADE[21] (Joint approximate diagonalization of eigenmatrices), FastICA[22], OGWE[23] (optimized generalized weighted estimator). It is important to know that being based on different independence criteria, each algorithm may lead to slightly different results.

In fact, every multidimensional recorded signal which can be considered like a combination of primordial independent signals (sources) can be treated by ICA, which tries to extract these independent sources and thus estimates the linear combination, i.e. a mixing matrix, who led to the observed multidimensional signal. Consequently, recorded signals in LC can be regarded as the sum of independent signals that constitute a DAD-chromatogram (i.e. peaks, noises, baseline

drift …). The use of UV-DAD detection and thus the hyphenated recorded signal make it possible to consider BSS of such chromatograms. It is thus very interesting to be able to separate peaks from noises or drift, and co-eluted peaks can be numerically separated for further processes. To carry out the numerical treatment in an automatic way, it is advisable to find a methodology which allows the use of ICA in a generic manner, i.e. to find the number of components/sources that ICA will try to make independent. In this paper, our objective is thus to investigate the possibility to perform automatically the treatment of data, ICA process, classification of components, peak picking, extraction of UV spectra for each component of a test mixture of different pharmaceutical compounds and reconstruction of cleaned DAD-chromatograms. This feasibility study on real data is the preliminary but very important step towards our final objective, the fully automated development of LC methods.

2. THEORY

2.1 Independent component analysis

Different definitions can be given for the Independent Component Analysis (ICA). One of the most classical is given below. This definition is referred as noise-free ICA. More details can be found in Hyvärinen and Oja's work[22,24] and related work of Hyvärinen.[25]

ICA of a vector **x** of signals values $(x_1,\ldots,x_m)^T$ consists of estimating the following generative model for the data:

$$\mathbf{x} = \mathbf{A}\mathbf{s}, \qquad (1)$$

where **A** is a constant ($m \times n$) mixing matrix, with m the number of observed signals, and the n components s_i in the signals vector $\mathbf{s}=(s_1,\ldots,s_n)^T$ are assumed independent. Both **A** and **s** must be estimated and n, the number of sources to be computed, must be chosen.

To solve this problem, the ICA algorithm estimates an unmixing matrix **W** such that the elements of the vector **s** are as statistically independent as possible. If **A** is a square matrix ($n=m$, i.e. ICA estimates as many independent components as the number of original recorded signals), the matrix **W** should be the inverse of the original mixing system **A** : $\mathbf{W}=\mathbf{A}^{-1}$. Notice that the ICA algorithms often allows to estimate less sources ($n<<m$) in order to simplify the computations. In this case, **A** is not square any longer. Generally speaking, the knowledge of an estimation of **W** is similar to the knowledge of the (estimated) mixing process described by **A**.

Then, after the ICA computation, it is possible to get the estimated sources:

$$\hat{\mathbf{W}}\mathbf{x} = \hat{\mathbf{W}}(\hat{\mathbf{A}}\hat{\mathbf{s}}) = (\hat{\mathbf{A}}^{-1}\hat{\mathbf{A}})\hat{\mathbf{s}} = \hat{\mathbf{s}}\,. \qquad (2)$$

An important step is to normalize sources to avoid infinity of solutions, when estimating **W**. The scaled factors for the sources is then contained in the **A** matrix. The j^{th} component is then defined as the product between the j^{th} line of the estimated A and the j^{th} element of the estimated s, and is formatted as x.

It should be noted that the statistical independence is a much stronger condition than non-correlation. Indeed, some uncorrelated functions are not necessarily independent. The independence between two sources s_1 and s_2 is estimated by:

$$E\{h_1(s_1)h_2(s_2)\} - E\{h_1(s_1)\}E\{h_2(s_2)\}. \qquad (3)$$

This expression will be equal to zero in case of independence, whatever the function h_1 and h_2.

The general hypotheses that must be assumed are listed below:

1. Signals and sources have *zero mean*.
2. Sources are assumed to be *statistically independent*
3. Sources must have unknown but *non-Gaussian* distributions (except one source).

However, ICA has two limitations. First, it only allows reconstructing non-Gaussian independent components (except one). Second, neither the signs, nor the order of independent components (i.e. the order of the lines of the **A** matrix and of the elements of the **s** vector) can be estimated. Fortunately, as can be seen below, these limitations are not a real problem for our purpose.

In practice, signals (**x** and **s**) are discretized and expressed as matrices, as explained in the next subsection.

2.2 ICA applied to HPLC-UV-DAD data

An UV-Diode-Array Detector (DAD) chromatogram can be considered as a matrix **X** containing, for each recorded time (the number of recorded time is *T*), the absorbance values at certain wavelengths *j*, in the range of the UV detector. **X** is then a matrix (*m* x *T*).

Thus, we dispose of *m* vector $x_j(t)$, representing the absorbance values at wavelength *j* as a function of the time (indexed by *t*). However, this definition is not in perfect accordance with the definition of ICA presented in the previous section. The matrix **X** is not precisely the presented vector. Fortunately, without loss of generality, one can suppose that the time axis of each $x_j(t)$ corresponds to different observations of a signal x_j. Thus, the **X** matrix can simply be expressed by the column vector $\mathbf{x}=(x_1,\ldots,x_j,\ldots,x_m)^T$, allowing the application of the ICA theory.

The hypotheses presented in the previous section are directly applicable. The first hypothesis can be easily fulfilled by subtracting the signal mean from each signal (chromatogram). The second hypothesis is the strongest one. However it is not unrealistic to assume that the peaks in a chromatogram are independent events, even if, from a chemical point of view, the related compounds may have similar structural features. This gives the ICA its power in the separation problem. Moreover, it does not need to be perfectly respected in practice. The third hypothesis is given for practical computational purpose.

One major concern in ICA is to find the sources that are relevant from those that are not. ICA has already been used in order to eliminate artifacts from data in electroencephalograms (EEG).[14,26] A similar methodology that the one presented by Delorme *et al.*[14] will be applied, with some extensions to automate the process.

Finally, related to this problem, a second concern is to determine the optimal number of sources *n* that must be used to break up the original DAD signal. The following sections present a new methodology to find this number of sources, applied in the context of design of experiments. In this framework, The DAD-chromatograms are referenced with the index *p*, (p=1,…,P).

2.3 High-order moments and statistics.

Since noise usually follows a Normal distribution, the computed independent sources given by ICA are investigated, no matter which value is chosen for *n*. It has been described in the literature that kurtosis value of a distribution is a good evidence to reject artifacts and noises[16]. This approach using kurtosis and other statistics or moments of the sources was used in the present study.

Thus, the following statistics or moments can be used to check the Normality of the distribution of each source. If normality is not observed, it should imply that the source is unlikely to be noise

and therefore likely corresponds to an exogenous phenomenon like a compound being detected. In summary, sources of interest such as peaks are assumed to have a non-normal distribution, and will be identified as such.

Kurtosis. The kurtosis is the fourth standardized moment of a random variable and is a measure of the peakedness of the probability distribution of this variable, and can be estimate as:

$$K_i = \frac{m_4}{m_2^2} - 3 = \frac{n\sum_{t=1}^{T}(\mathrm{s}_i(t) - \bar{\mathrm{s}}_i)^4}{\left(\sum_{t=1}^{T}(\mathrm{s}_i(t) - \bar{\mathrm{s}}_i)^2\right)^2} - 3 \qquad \textbf{(4)}$$

m_k represents the moment (about the mean) of order k and $\bar{\mathrm{s}}_i$ is the mean of the i^{th} source s_i.

Shapiro-Wilk statistics. The non parametric Shapiro-Wilk (S-W) test tests the null hypothesis that a sample $\mathrm{s}_i(1), ..., \mathrm{s}_i(T)$ comes from a Normally distributed population. This hypothesis is rejected if the test statistics is too small. This test statistic or the associated *p*-value can be used to characterize the distribution. More details about this test can be found in the literature[27].

Kurtosis and S-W statistics can be used to describe the degree of Normality of a distribution and, hence, to identify noise sources. However, this is far to be the only statistics allowing this, and other simple or complex statistics (skewness of the distribution of the component, range, Kolmogorov-Smirnov Normality test, etc.) can also be elegantly used in this context.

2.4 *k*-means clustering

k-means is a non supervised method to cluster objects into *k* partitions on the basis of their attributes. The objects are the sources computed by ICA and the attributes are the estimated moments and statistics computed on these sources. The aim is to use these computed characteristics to discriminate the relevant sources (peaks) from the noise or irrelevant artifacts, chromatogram by chromatogram. The concept is to compare the Euclidian distance between the objects. A short distance (slight difference on the computed attributes) is a sign of closeness between objects. The closest objects are putted together in the same cluster.

Different implementations exist for *k*-means clustering. The Hartigan and Wong one was used here, because it generally does a better job than other implementations[28]. However our application of clustering is rather simple, with a low number of variables and observations, so that the way of implementation is not very relevant.

Before clustering is applied, attributes are standardized by dividing by their standard deviation (computed on the total number of sources available on the DAD-chromatogram). This is done to give each attributes the same weight in the decision process. This is illustrated by the following equation, applied to kurtosis. The same computation was done for each considered statistics.

$$StandarizedKurtosis_i = \frac{K_i}{\sqrt{\sum_{i=1}^{n}(K_i - \bar{K})^2}}, \qquad \textbf{(5)}$$

where $\bar{K}$ is the mean of the kurtosis computed on all the sources of one chromatogram.

In practice, *k*-means with *k*=2 is used. Figure 1 shows the expected positions of both clusters, when *k*-means is applied on the attributes kurtosis and S-W statistics. The partitioning should result in a cluster with sources having a high Kurtosis and a low S-W values for the sources of interest (cluster 1) and conversely a low Kurtosis and a high S-W values for noises and artifacts

sources (cluster 2). For simplicity, clustering algorithm is assumed to be determinist (always giving the same results on the same set of data), although it is not the case. For the following sections, the number of sources in the relevant cluster for the p^{th} DAD-chromatogram is defined as $n_r(n,p)$, as it also depends of n.

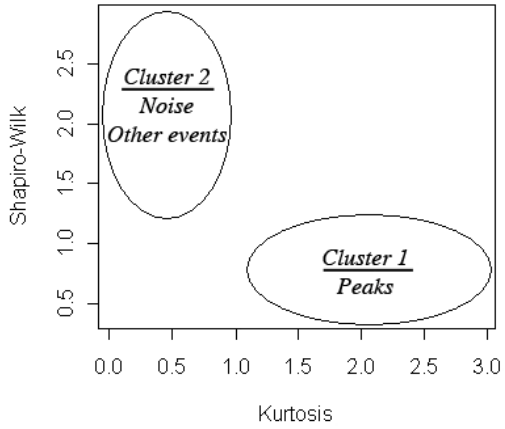

Figure 1. Expected clustering of sources using normalized high-order statistics (kurtosis and S-W statistics are used) for one DAD-chromatogram.

2.5 Selection of the number of sources for computation

As previously explained, the main parameter of the ICA is the number n of sources or components chosen to separate the signal. It is by default arbitrarily chosen by the operator but, for our purpose, an automated process was developed, that will find the optimal number of sources to be estimated ignoring the expected number of peaks to be detected. As our approach is created to be flexible and should have the ability to find unexpected peaks, it has to be done by a step-by-step automated process.

Yuzhu *et al.* showed that the calculation of singular value ratios (SVR) is useful to estimate the number of "pure" components in HPLC-DAD data[29], i.e. an estimation of the number of peaks. This number can be used to determine the number of sources needed to separate these pure components. Other techniques, such as the squared difference between an original and reconstructed signal, are available but are not usable in our context[26]. The square difference would lead to reject components corresponding to a very small, but relevant, peak, because it does not significantly contribute to decrease the sum of squares of errors (SSE).

Our approach is different because we want to determine this number using ICA and clustering methods in a DOE framework. In DOE, the same mixture is injected several times using different analytical conditions. Thus, each DAD-chromatogram should contain the same information, i.e. the same number of peaks/artifacts (unless some impurities appear). Finding the number n of sources for the application of ICA is similar as finding the most likely value of $n_r(n,p)$ across all DAD-chromatograms. This most likely value is defined as $n_r^*(n)$. For one value of n, the Figure 2 illustrates how to locate $n_r^*(n)$, the expected number of peaks to extract out of n sources across the P DAD-chromatograms.

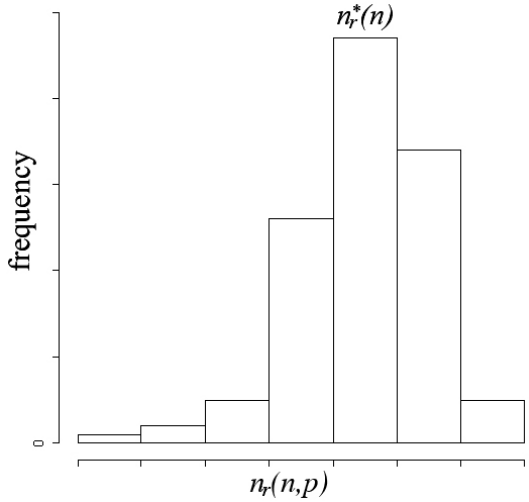

Figure 2. Example of optimal value of $n_r(n,p)$, computed across the P DAD-chromatograms, for a specific value of n.

This innovative strategy is presented in Figure 3. An initial value for n must first be given. One can use an estimated number of pure components, with SVR computation, for instance. Another easier technique is to begin the ICA computation with a very low number of sources, say n=3. Then, ICA is performed with this initial value of n. For each estimated sources, the moments and statistics presented in the previous sections are computed, and the k-means clustering analysis (with k=2) is applied on these (standardized) values. The relevant cluster contains $n_r(n,p)$ sources identified as peaks or important artifacts. This process is repeated, incrementing value of n for the P DAD-chromatograms, and then each value of $n_r(n,p)$ is recorded. It is possible to realize a plot of $n_r^*(n)$ against the values of n (Figure 8). One can observe that $n_r^*(n)$ stabilizes when n increases. This stabilized value, n_c^*, will be assumed to be the optimal number of peaks and relevant artifacts in the mixture.

The final step is to ensure that n_c^* relevant peaks will be found for all DAD-chromatograms. This may be done by investigating in the recorded results, for each DAD-chromatogram, how many sources (n) were necessary to get n_c^* components in the good cluster.

The main disadvantage of this technique is the computational time, when working on DAD-chromatograms issued from a design with many experiments (e.g. 50) and when the number of sources n has to be increased because many compounds are concerned. The estimation of the moments and statistics, clustering application and other processes, realized on the whole set of components for each chromatogram can become quite long. However, the gain is that no manual work has to be done to find peaks in noisy DAD-chromatograms.

3. MATERIALS AND METHODS

3.1 Chemicals.

Methanol and acetonitrile were HPLC grade from Sigma (St-Louis, MO, USA). Ultra pure water was obtained with a Millipore (Billerica, MA, USA) Milli-Q Academic A10. Atenolol, phenytoine, sulfinpyrazone and warfarin were used as reference compounds throughout the study. These pharmaceutical compounds were selected for the difference existing between their chemical compositions, the large disparity between some of their physico-chemical descriptors (e.g. LogP - pKa) leads to different selectivity among the obtained chromatograms. These four compounds were obtained as reference from the Eli Lilly Company (Indianapolis, IN, USA).

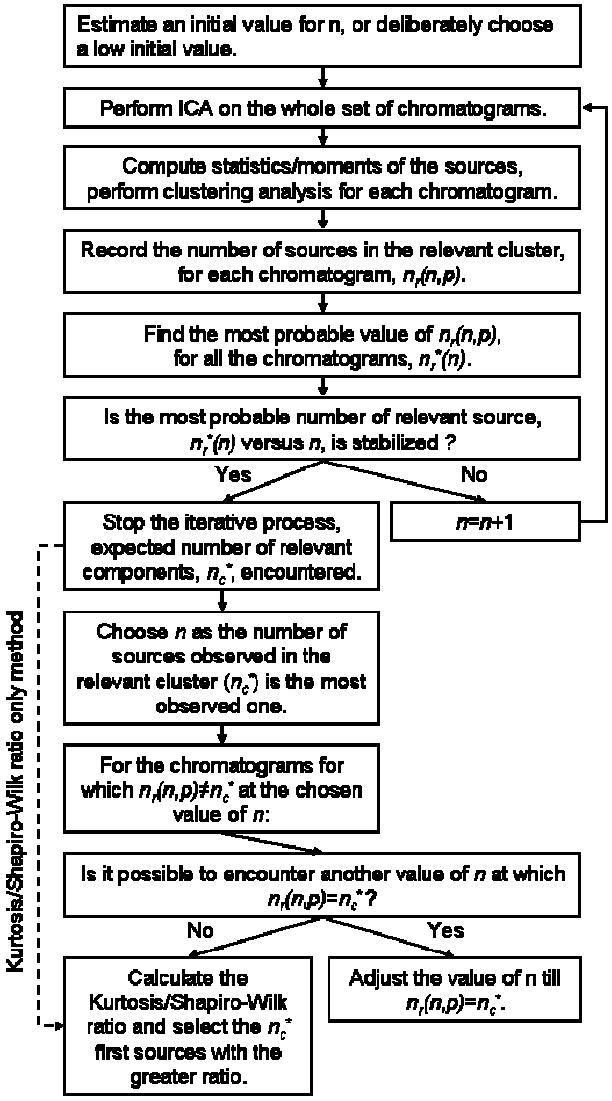

Figure 3. Methodology for the selection of number of sources for the ICA computation.

3.2 Experiments.

A full factorial design[30] has been applied on four factors: stationary phase, pH of the aqueous part of the mobile phase, gradient slope and the nature of organic modifier. Five analytical columns were used: C18, C8, RP18, Phenyl XBridge columns (100x2.1 mm i.d.; particle size 3.5 μm), all from Waters (Milford, MA, USA) and a C18 Cogent Bidentate column (100x2.1 mm i.d.; particle size 4.0 μm) from Microsolv (Eatontown, NJ, USA). The experiments were carried out at a flow rate of 0.25 mL/min and at 30 °C. The buffers consisted in 10 mM of concentrated formic acid for pH 2.6, ammonium formate for pH 5.0 and pH 7.0 and ammonium hydrogencarbonate for pH 10.0. The pH was adjusted to the desired value with concentrated formic acid or ammonia 35% aqueous solution. The shapes of the linear gradients are described in Table 1.

Chromatographic separation were performed on a Waters 2695 separation module coupled to a Waters selector valve 7678 and a Waters 996 Photodiode array detector. All DAD-chromatograms were recorded between 210 nm and 394 nm with an estimated step of 1.2 nm (158 points) and with a time resolution of 500 ms. They were finally exported by Empower 1.0 (Waters) in an ASCII file containing the UV-DAD matrix.

The experimental design leads to 5·2·4·3 = 120 experiments. The previously described methodology has been then applied on the resulting 120 UV-DAD chromatograms.

Time (min.)	O.M.%	Time (min.)	O.M.%	Time (min.)	O.M.%
0	5	0	5	0	5
10	95	20	95	30	95
10.5	95	20.5	95	30.5	95
10.6	5	20.6	5	30.6	5
16	5	26	5	36	5

Table 1. Linear gradient shapes with organic modifier percentage (O.M.%) in mobile phase.

3.3 Software.

An in-house computer program was developed to perform the analysis presented in the previous sections. The coding was carried out with R 2.4.1 statistical language for Windows, freely distributed at http://www.r-project.org. These codes can be run on compatible PC or other environments where R is available. For ICA computations, FastICA algorithm for R, developed at the Helsinki University of Technology, was used.

4. RESULTS AND DISCUSSION

4.1 UV-DAD Data.

The ICA approach is applied on each DAD-chromatogram. Each **X** chromatogram matrix can be described as a (158 x *T*) matrix where *T* is either 1920, 3120 or 4320, depending in the analytical conditions of gradient (10, 20 and 30 min, respectively).

The methodology to find *n* and $n_r(n,p)$ is illustrated with *n*=12 on the DAD-chromatogram recorded on the XBridge C18 column with methanol and pH 5 buffer, with a gradient of 20 minutes (Figure 4). In fact *n*=12 is a good value for *n* because it succeeds in finding the 4 compounds and the dead volume perturbation (i.e. the small baseline perturbation observed at t_0) of the considered DAD-chromatogram. The way to find systematically the value of *n* for each DAD-chromatogram of the design is presented.

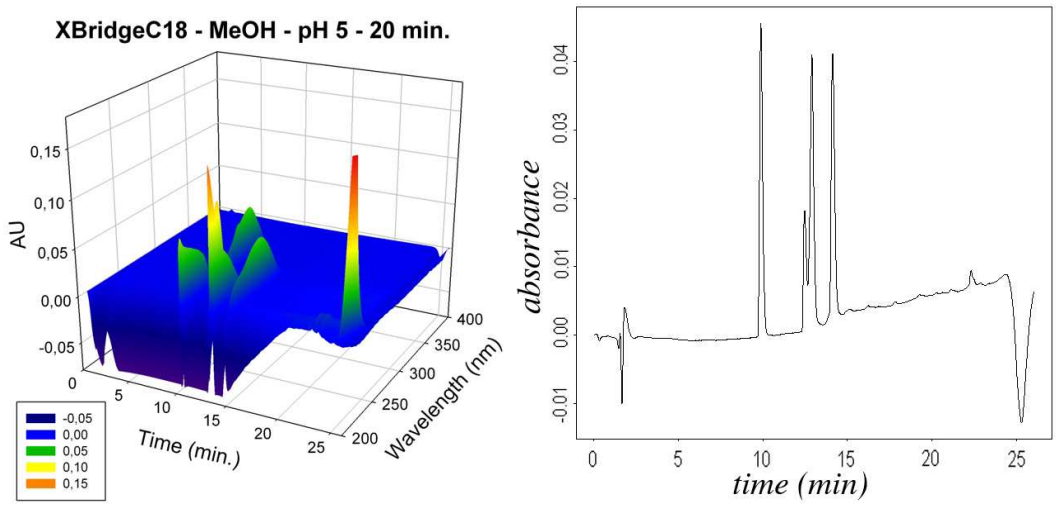

Figure 4. (left) Initial DAD-chromatogram recorded on an XBridge C18 with methanol and pH 5 buffer with a gradient of 20 minutes. (right) DAD-chromatogram observed at 240 nm.

A first data cleaning has consisted in the cut of each DAD-chromatogram at 14, 24, 34 min for gradient times of 10, 20, 30 min respectively. This was performed in order to remove the perturbation of the end of the gradient, containing only noise or irrelevant artifacts.

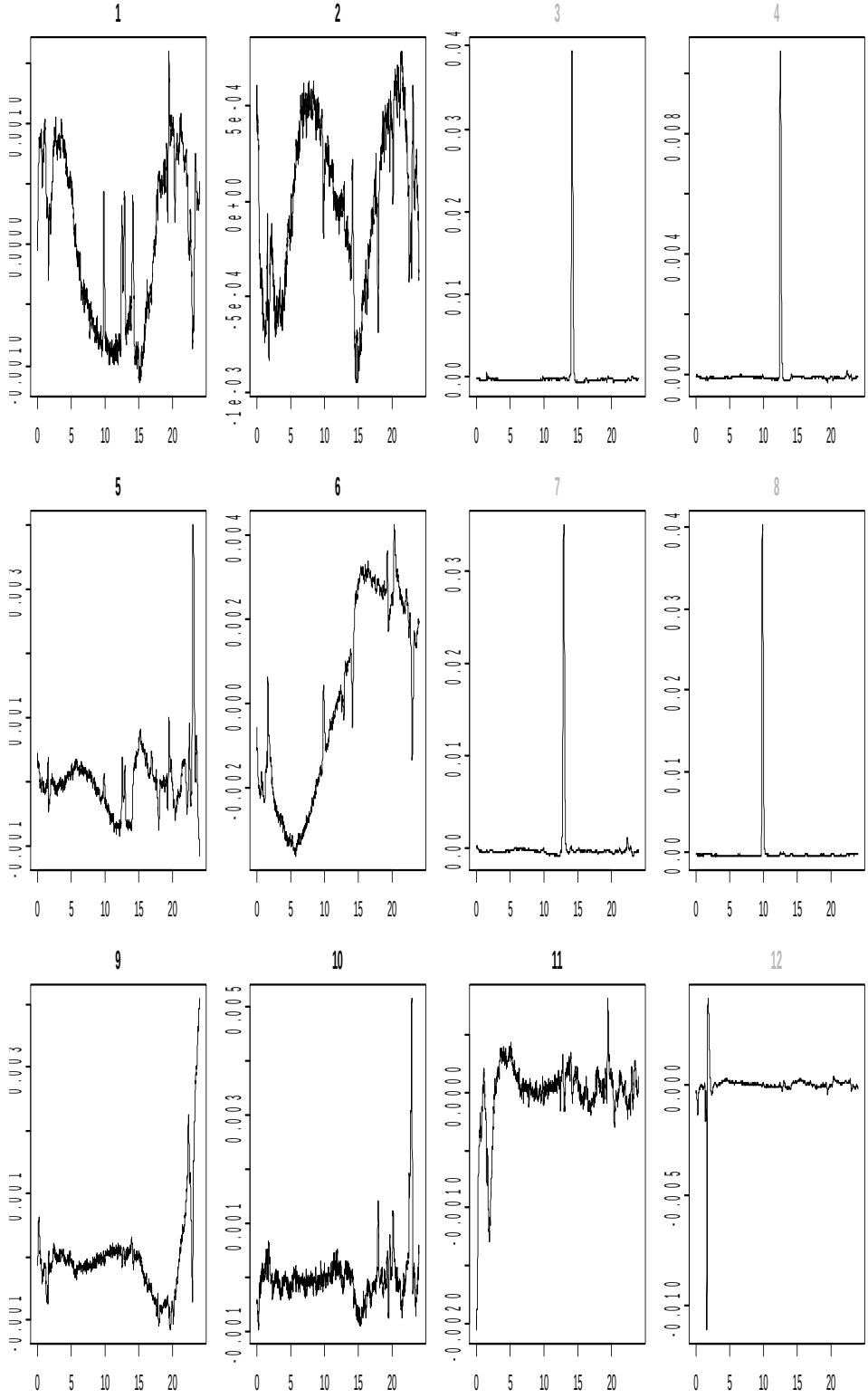

Figure 5. Components computed by ICA with *n*=12 observed at 240 nm. The components numbered in grey are the ones that we can observe they correspond to peaks or relevant artifact. Components marked in black are noise. The X-axis represents the time (min.) and the Y-axis is the absorbance of the components.

ICA was then applied on the DAD-chromatogram, resulting in 12 components, as shown in Figure 5. Each independent component at one wavelength or a UV spectrum at a given time was then displayed. Figure 5 presents an example of this display at the wavelength of 240 nm.

The components corresponding to peaks and relevant artifacts are voluntary shown (sources 3,4,7,8 and 12). At the opposite, the seven other components labeled in black seemingly correspond to noise. The aim was to automatically detect the relevant components.

4.2 Clustering and optimal number of sources.

The kurtosis and the Shapiro-Wilk statistics were computed on each source. After standardization, *k*-means clustering was applied. The results are shown in Figure 6 (a). The cross located at the right and bottom indicates the center of the relevant cluster (cluster 1). The sources-components contained in this cluster are effectively the ones that correspond to peaks and relevant artifacts. The maximum of each relevant component (apex) has been found and its retention time has been automatically placed on the chromatogram (Figure 6, b).

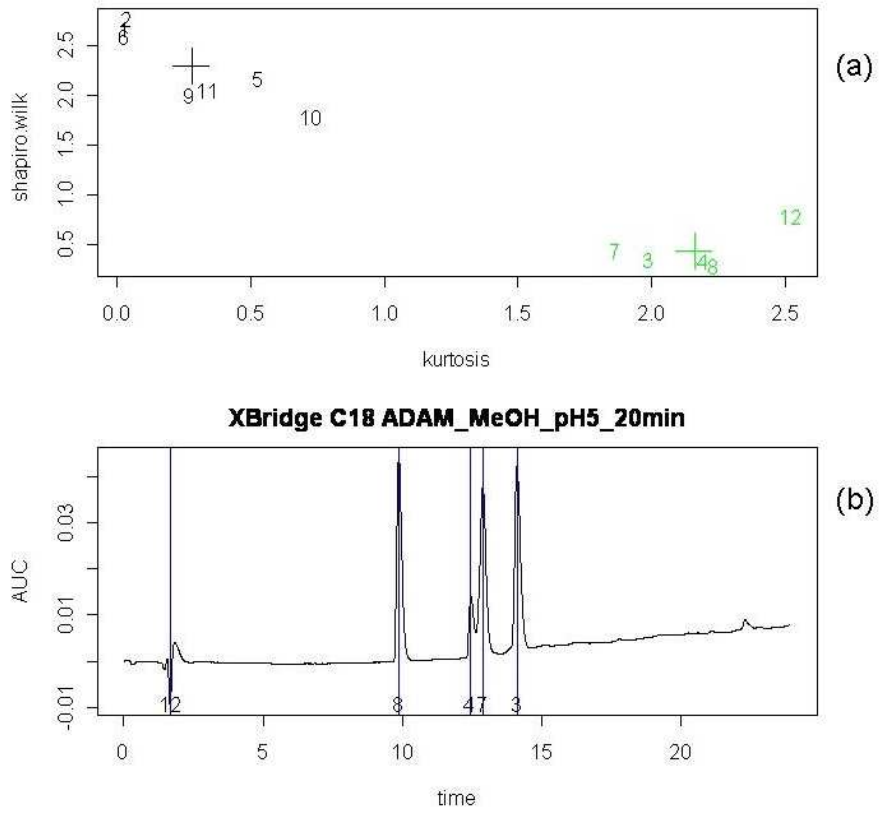

Figure 6. (a) Clustering *k*-means realized on standardized kurtosis and Shapiro-Wilk statistics. The right-bottom cluster (cluster 1, in grey) contains the relevant sources. Crosses are the clusters centers. (b) Original DAD-chromatogram at 240 nm with automatically picked apexes by a vertical line.

With *n*=12, 40.0% of results counted five components in the relevant cluster, across all the 120 DAD-chromatograms. The distribution of the number of components counted in the relevant cluster, for *n*=12, is displayed in Figure 7.

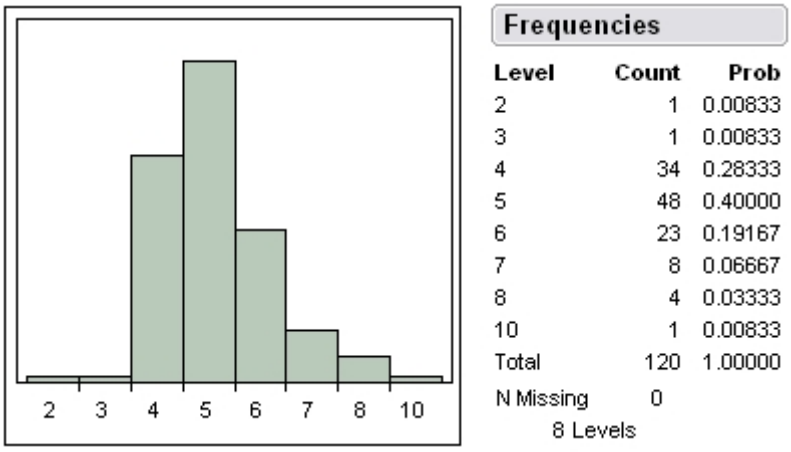

Figure 7. Distribution of $n_r^*(n)$ for the 120 DAD-chromatograms with n=12.

The same process has been applied for each value of *n.* The ICA process was restarted while changing *n* and observing how this percentage was affected. *n* was incremented from 3 to 30 and the result is illustrated on the Figure 8. It shows a stabilization of number of relevant components $n_r^*(n)$ at 5 components, for n=12 and upper. So, n_c^* can be assumed equal to 5.

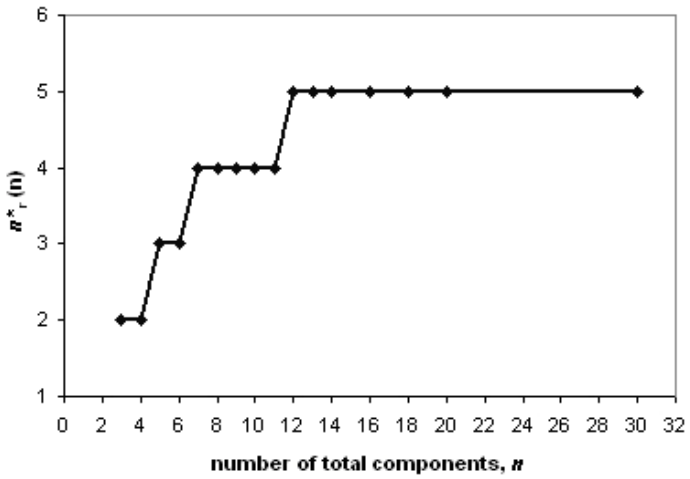

Figure 8. Plot of the variation of $n_r^*(n)$ versus *n*. $n_r^*(n)$ stabilizes at the value of n_c^*=5.

Five relevant sources were thus preferably counted in the relevant cluster. This result indicates that each sample seems to contain five independent components of interest. This conclusion is correct because the four injected compounds and the dead volume perturbation of the baseline can be found.

4.3 Dealing with results.

However, for n=12, $n_r(n,p)$ differs from n_c^*=5 in 60.0% of the clustering results. Then, for these inconvenient results, it is proposed to adapt the value of *n* till $n_r(n,p)$ reaches n_c^*, investigating the recorded results. However in some cases no value of *n* makes it possible to reach this objective. So, the first n_c^* *best* sources were defined as being the relevant ones. It is carried out by calculating a ranking index, the $\frac{\text{Kurtosis}}{\text{Shapiro - Wilk}}$ ratio for each source and to keep those with the largest ratio.

In order to remain consistent, this approach was also tested on the whole set of DAD-chromatograms. Results of both automated peak picking approaches are exposed in Table 2.

4.4 Extracting sources UV-spectra.

Figure 9 shows that an estimated version of the UV spectrum (b) for each relevant peak (a) can be recovered and compared to the reference UV spectra of the investigated compounds (c), if known. This can be of great help for the identification of the compounds and would even be used to automatically carry out peaks' matching. However some problems can occur. First, it is mandatory to get estimations of the UV spectra that are precise enough. Nevertheless, co-elution often distorts estimated concerned UV spectra. Second, The UV spectrum of certain compounds is logically modified by the buffer pH. For these reasons, automatic peaks' matching remains a matter of particular interest.

4.5 Reconstruction of DAD-chromatogram.

Once the relevant sources were automatically found, the reconstruction process could be started by simply summing all the selected components, allowing retrieving a clean DAD-chromatogram. Figure 10 (a) and (c) presents the original DAD-chromatogram compared to the reconstructed one, both observed at 240 nm. The peaks (identified by the triangles in the top of the chromatograms) have been automatically picked. However, at this wavelength, the purification of the chromatogram is not really observable. The chromatograms of Figure 10 (b) and (d) illustrate that, for every wavelengths, the reconstructed chromatogram (d) has been cleaned from artifact or gradient effects.

4.6 Results of automated peak picking.

Finally, the automated "peak picking study" was compared to the hand-made one. Results are reported in Table 2. 95.8 % of the four automatically picked peaks correspond to the four manually picked peaks. However, even done by a confirmed analyst, the manual peaks' picking can not always be error free (but we assumed it is error free for these values). The dead volume perturbation is of particular interest (defining the dead time of a HPLC system), but the picking of this artifact is far not as good. This is due to some impurities or very significant artifacts, whose sources look as relevant one.

Clustering Method	*picked peaks*	*non picked peaks*		*picked relevant artifacts*	*Picked irrelevant artifacts or impurities*
*Count (total:120*4)*	457	23	*Count (total = 120)*	59	49
%	95.2 %	4.8 %			
Kurtosis/ Shapiro-Wilk ratio method	*picked peaks*	*non picked peaks*		*picked relevant artifacts*	*picked irrelevant artifacts or impurities*
*Count (total:120*4)*	460	20	*Count (total = 120)*	59	71
%	95.8 %	3.5 %			

Table 2. Comparison of automated peak picking results on the complete set of chromatograms by the two methods.

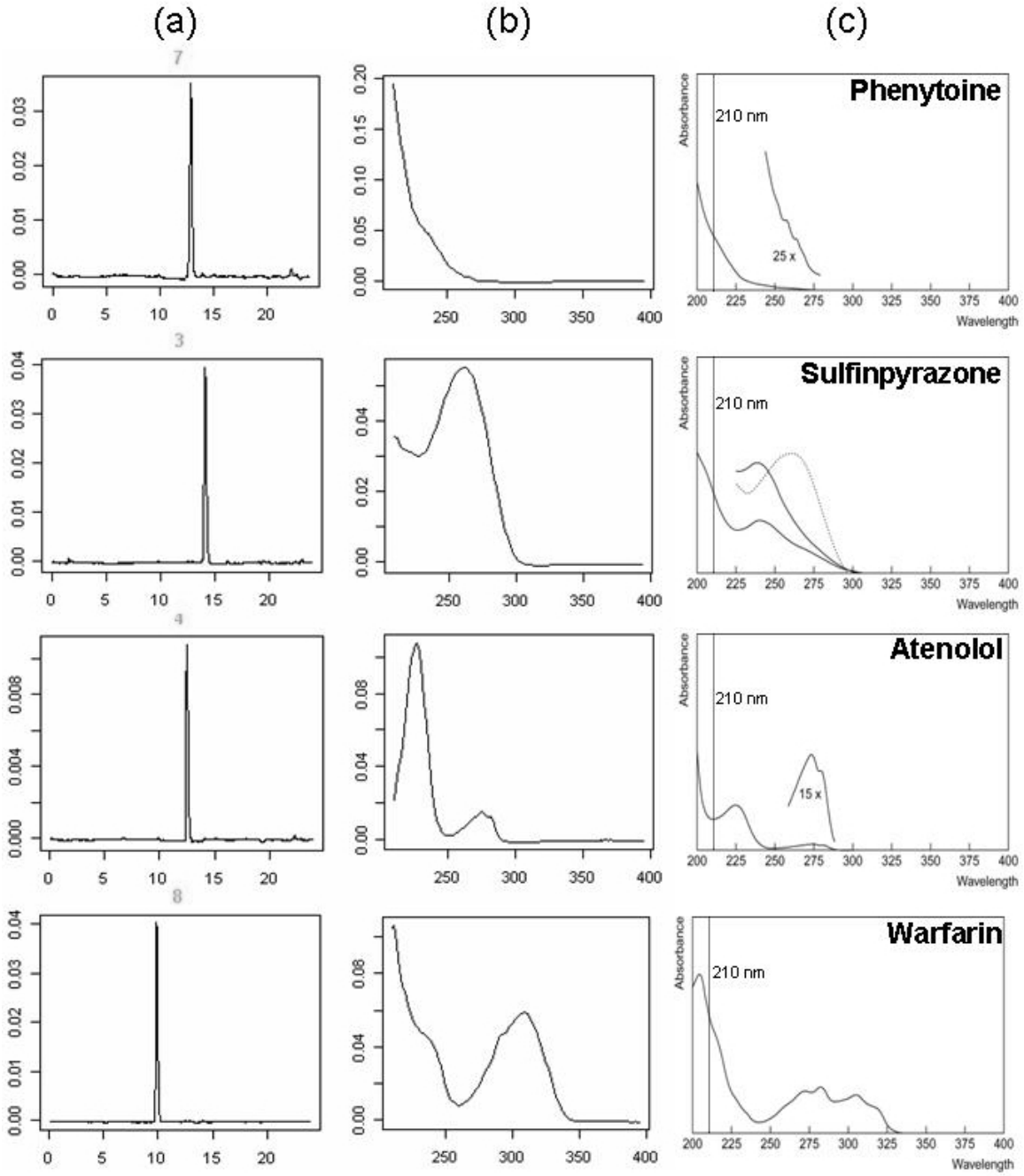

Figure 9. Obtained relevant components and comparison between their UV-spectra and reference ones. (a) Components contained in the relevant cluster, observed at 240nm; (b) Estimated UV spectra of each independent component at apex (210-400 nm); (c) Reference UV spectra (200-400 nm) from Clarke 2004[31]. The reference spectra recorded in basic media are shown in dotted line.

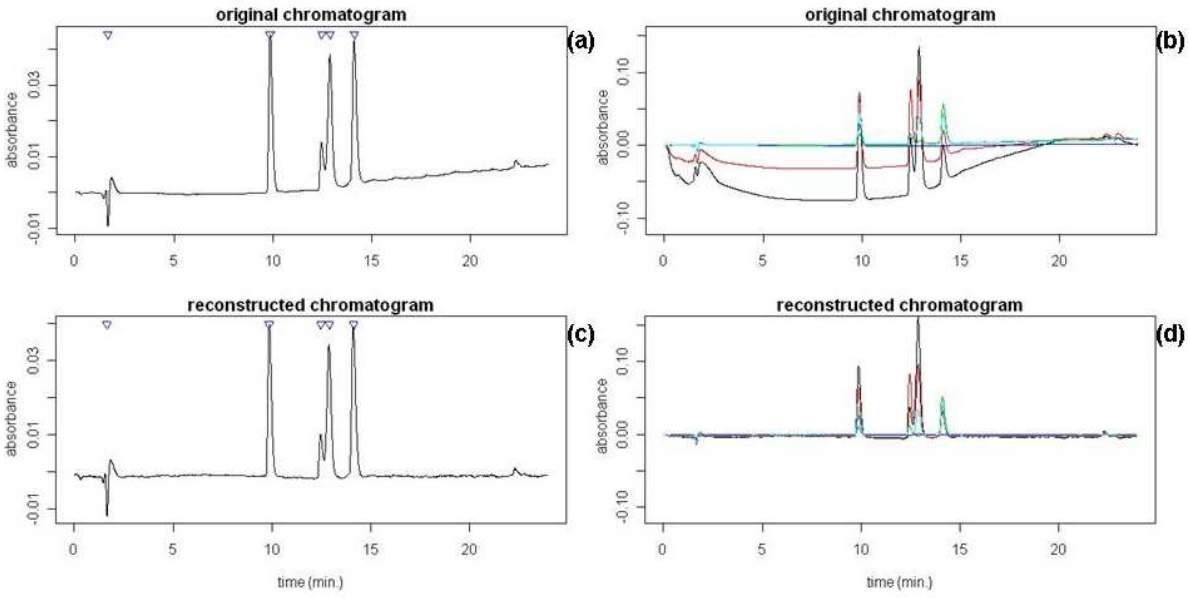

Figure 10. (a-c) Comparison between original DAD-chromatogram (a) and reconstructed one with *n*=12 (c). The original is recorded at 240 nm, on an XBridge C18 with methanol and pH 5 buffer with a gradient of 20 minutes. Apexes have been automatically picked. (b-d) Comparison at different wavelengths (214, 220, 240, 268 and 327 nm) of the same original (b) and reconstructed (d) chromatograms.

5. CONCLUSIONS

The automated finding of peaks is a very crucial step in the automated development of analytical methods. This new and original approach combining design of experiments, ICA, high-order statistics and clustering is very powerful and promising as illustrated by its successfully application for the determination of a test mixture of pharmaceutical compounds. On the other hand, great advantages of the present approach would be found when dealing with high throughput screening experiments, as those resulting of the follow-up of the synthesis process (purity assessment) or in the framework of the development of stability indicating method where impurities are not necessary known. Moreover, it does not require expensive equipments, such as mass spectrometer, to detect all compounds of a sample. Clustering methods allow separating very efficiently the noise components from the relevant ones, using adequate summary statistics. The technique to find an optimal number of sources is very convenient but is computationally expensive. Fortunately, the time needed for the numerical data treatments presented in this study is smaller than the sample analysis time, and the chromatograms are numerically treated one by one. This gives the opportunity to easily implement the numerical data treatments in concurrent mode, while the sample analysis are processing.

Finally, this process could also be performed only on sub-parts of a DAD-chromatogram, according to the analyst interest; for instance, searching of co-eluted impurities in peaks of interest.

ACKNOWLEDGMENTS

The authors would like to thank the Walloon Region of Belgium for the FIRST-DEI convention funds N°516130 and the Eli Lilly Company for partial funding of ADAM project and equipments.

REFERENCES

[1] Torres-Lapasio, J.R.; Garcia-Alvarez-Coque, M.C. *J. Chromatogr., A* **2006,** *1120*, 308-321.

[2] Krupcik, J.; Mydlova, J.; Spanik, I.; Tienpont, B.; Sandra, P. *J. Chromatogr., A* **2004**, *1084*, 80-89.

[3] Vivó-Truyols, G., Torres-Lapasió, J.R, García-Alvarez-Coque, M.C. *J. Chromatogr., A* **2003**, *991*, 47-59.

[4] di Marco, V.B.; Bombi, G.G. *J. Chromatogr., A* **2001**, *931*, 1.

[5] Nikitas, P.; Pappa-Louisi, A.; Papageorgiou, A. *J. Chromatogr., A* **2001**, *912*, 13.

[6] Torres-Lapasio, J.R.; Baeza-Baeza, J.J.; Garcia-Alvarez-Coque, M.C. *Anal. Chem.* **1997**, *69*, 3822.

[7] Bogomolov, A.; McBrien, M. *Anal. Chim. Acta* **2003**, *490*, 41-58.

[8] Tauler, R. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 133-146.

[9] de Juan, A.; Tauler, R. *J. Chromatogr., A* **2007**, *1158*, 184-195.

[10] Boulanger, B.; Dewé, W.; Ceccato, A.; Debrus, B.; Lebrun, P.; Hubert, Ph. ICAChrom: Utilisation de l'Analyse en Composantes Indépendantes (ICA) pour la séparation numérique des pics et la quantification automatique en CLHP-UV, Chimiométrie 2006, Paris, France.

[11] Debrus, B.; Lebrun, P.; Boulanger, B.; Dewé, W.; Ceccato, A.; Hubert, Ph. Séparation numérique et quantification automatique de pics en CLHP-UV grâce à l'utilisation de l'analyse en composantes indépendantes (ICA), Sep 07, Grenoble, France.

[12] Lebrun, P.; Debrus, B.; Boulanger, B.; Caliaro, G.; Ceccato,A.; Hubert, Ph. Numerical separation in HPLC-UV-DAD with Independent Component Analysis (ICA) using high order statistic for the automated identification of peaks, HPLC 2007, Ghent, Belgium

[13] Makeig, S.; Jung, T.; Bell, A. J.; Ghahremani, D.; Sejnowski, T. J. *Proc. Natl. Acad. Sci. U. S. A.* **1997**, *94*, 10979-10984.

[14] Delorme, A.; Sejnowski, T.; Makeig, S. *NeuroImage* **2007**, *34(4)*, 1443-1449.

[15] Jung, T.-P.; Humphries, C.; Lee, T.-W.; Makeig, S.; McKeown, M.; Iragui., V.; Sejnowski, T. J. *Advances in Neural Information Processing Systems* **1998**, *10*, 105394-900.

[16] McKeown, M.; Makeig, S.; Brown, G.; Jung, T.-P.; Kindermann, S.; Lee, T. W.; Sejnowski, T. J. *Proc. Natl. Acad. Sci. U. S. A.* **1998**, *95*, 803-810.

[17] Yamamoto, H.; Hada, K.; Yamaji, H.; Katsuda, T.; Ohno, H.; Fukuda, H. *Biochem. Eng. J.* **2006**, *32*, 149-156.

[18] Wang, G.; Cai, W.; Shao, X. *Chemom. Intell. Lab. Syst.* **2006**, *82*, 137-144.

[19] Vosough, M. *Anal. Chim. Acta* **2007**, *598*, 219-226.

[20] Jouan-Rimbaud Bouveresse, D.; Benadid, H.; Rutledge, D.N. *Anal. Chim. Acta* **2007**, *589*, 216-224.

[21] Cardoso, J.-F., *Neural Comput.* **1999**, *11(1)*, 157.

[22] Hyvärinen, A.; Oja E. *Neural Comput.* **1999**, *9*, 1483.

[23] Murillo-Fuentes, J.J.; Boloix-Tortosa, R.; Gonzalez-Serrano, F.J. Fourth International Symposium on Independent Component Analysis and Blind Signal Separation (ICA2003), Nara, Japan, **2003**, pp 1053-1058.

[24] Hyvärinen, A.; Oja, E. *Neural Networks* **2000**, *13(4-5)*, 411-430

[25] Hyvärinen, A. *IEEE Trans. on Neural Networks*, **1999**, *10(3)*, 626-634.

[26] Bugli, C., *Statistical tools for the analysis of event-related potentials in electroencephalograms*, PhD thesis, Institut de statistiques, Université catholique de Louvain, Louvain-la-Neuve, **2006**.

[27] Shapiro, S. S.; Wilk, M. B. *Biometrika* **1965**, *52 (3-4)*, 591-611.

[28] Hartigan, J. A.; Wong, M. A. *Applied Statistics* **1979**, *28*, 100–108.

[29] Yuzhu, H.; Weiyang, S.; Weifeng, Y.; Massart, D.L. *Chemom. Intell. Lab. Syst.*, **2005**, *77(1-2)*, 97-103.

[30] Cox, G.; Cochran, W. *Experimental Designs*, 2nd ed., McGraw-Hill, 1962.

[31] Moffat, A. C.; Osselton, M. D.; Widdop, B. (eds), *Clarke's Analysis of Drugs and Poisons*, Pharmaceutical Press, Electronic version, (Edition 2004).