Two novel methods for the determination of the number of components in independent components analysis models

D. Jouan-Rimbaud Bouveresse [a,b,*], A. Moya-González [c], F. Ammari [d], D.N. Rutledge [a,b]

[a] INRA, UMR 1145 Ingénierie Procédés Aliments, F-75005, Paris, France
[b] AgroParisTech, UMR 1145 Ingénierie Procédés Aliments, F-75005 Paris, France
[c] LPF-Tagralia, Universidad Politécnica de Madrid, Madrid, Spain
[d] Faculté des Sciences de Bizerte Jarzouna-7021, Université 7 Novembre Carthage, Tunis, Tunisia

Keywords:
Independent Components Analysis (ICA)
Validation
Durbin–Watson criterion

A B S T R A C T

Independent Components Analysis is a Blind Source Separation method that aims to find the pure source signals mixed together in unknown proportions in the observed signals under study. It does this by searching for factors which are mutually statistically independent. It can thus be classified among the latent-variable based methods. Like other methods based on latent variables, a careful investigation has to be carried out to find out which factors are significant and which are not. Therefore, it is important to dispose of a validation procedure to decide on the optimal number of independent components to include in the final model. This can be made complicated by the fact that two consecutive models may differ in the order and signs of similarly-indexed ICs. As well, the structure of the extracted sources can change as a function of the number of factors calculated. Two methods for determining the optimal number of ICs are proposed in this article and applied to simulated and real datasets to demonstrate their performance.

1. Introduction

Independent Components Analysis (ICA) [1,2] is a method that was developed in the early 1980s in the domain of signal processing. It is now being applied to all domains where signals have to be analysed, including analytical chemistry [3], but also in statistical process control [4], in the biomedical field (e.g., for the removal of ocular artefacts from electroencephalograms [5,6], or for the detection of brain tumours [7,8]) and in image analysis [9]. ICA can be used to remove artefacts [7,8,10,11], separate sources [12–14] or identify constituents in a mixture [14–17].

ICA is a latent variable-based method, i.e. it is based on the construction of latent variables, or factors, called *Independent Components* (ICs), which are linear combinations of the original variables. The ICs are assumed to correspond to the signals of the pure sources present in the analysed mixtures (or to signals related to the pure sources, as in ref. [18] for instance, where the ICs correspond to the difference spectra of the pure components taken two by two). The hypothesis used to enable the extraction of the "pure source signals" is that these vectors are statistically independent, as opposed to Principal Components Analysis (PCA) [19] which is based on calculating orthogonal vectors that maximise the amount of variance extracted from the data. As there are several approaches in assessing statistical independence, there exist several different ICA algorithms (FastICA [20], Joint Approximate Diagonalization of Eigenmatrices (JADE) [21], InfoMax [22], Mean-Field ICA [23], Kernel ICA [24,25], Mutual Information Least Dependent Component Analysis [26,27], Stochastic Non-Negative Independent Components Analysis [27,28], Robust Accurate Direct Independent Components Analysis aLgorithm (RADICAL) [29]).

One major issue when using a model based on latent variables (LVs) is the determination of the number of LVs to use. This can be done by building A models ($A>1$), with from 1 to A factors respectively, and by estimating the error (or another statistic) associated with each model. When an optimum is reached, or when no significant change in the statistic used is observed, the corresponding number of LVs is considered to be optimal. The validation of models performed in this way presupposes that when two consecutive models are constructed, the corresponding LVs are related to the same information in the data. This is the case for nested models (such as PCA [19], Principal Components Regression (PCR) [30], or Partial Least-Squares (PLS) regression [31], for example), where the $A+1$-factor model corresponds to the A-factor model to which factor $A+1$ is added. ICA models, however, are not nested, and it can occur that when going from the A-factor model to the $A+1$-factor model, some ICs do not have the same index in both models, and/or do not have the same sign, or are simply different.

This problem can also arise if standard cross-validation (CV) [32,33] is used to determine the optimal dimensionality of an ICA model: it may happen that the set of A ICs built after removal of cross-validation segment n are different from the set of ICs built after

* Corresponding author at: INRA, UMR 1145 Ingénierie Procédés Aliments, F-75005, Paris, France. Tel.: +33 1 44 08 16 39; fax: +33 1 44 08 16 53.
E-mail address: delphine.bouveresse@agroparistech.fr (D. Jouan-Rimbaud Bouveresse).

removal of cross-validation segment m ($m \neq n$). Hence, precautions need to be taken when cross-validating ICA models. The use of Procrustes rotation proposed by Westad et al. [18] partly solves this ambiguity. In order to compare the quality of consecutive ICA models, and help choose the appropriate number of ICs to include in the model, an uncertainty parameter was also proposed, defined as in Eq. (1):

$$s^2(s_a) = \left(\sum_{m=1}^{M} \left(s_a - s_{a(-m)}\right)^2\right)\left(\frac{M-1}{M}\right) \quad (1)$$

where $s^2(\boldsymbol{s}_a)$ is the estimated uncertainty variance of the ath ICA loading, M is the number of cross-validation segments, $\boldsymbol{s}_a$ is the ath loading vector of the ICA model built with all objects, while $\boldsymbol{s}_a(-m)$ corresponds to the ath loading vector from the model built after removal of cross-validation segment m.

Wang et al. [34] have also proposed to build consecutive ICA models (with from 1 to A ICs), to reconstruct the **X** matrix with each of them, yielding $\boldsymbol{X}_a$ when the a-IC ICA model is built ($1 \leq a \leq A$), and to compute a Sum of Squared Residuals (SSR) between the original and the reconstructed **X** for each model. The model corresponding to a minimal SSR is optimal:

$$\mathrm{SSR} = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n}\left(X_{ij} - \hat{X}_{ij}\right)^2} \quad (2)$$

Another interesting parameter to estimate the quality of an ICA model in the Amari index [35], defined as:

$$P_{err} = \frac{1}{2N}\sum_{i,j=1}^{N}\left(\frac{|p_{ij}|}{\max_k(|p_{ik}|)} + \frac{|p_{ij}|}{\max_k(|p_{kj}|)}\right) - 1 \quad (3)$$

where $p_{ij} = (\hat{\mathbf{A}}^{-1}\mathbf{A})_{ij}$, the matrix **A** being the mixing matrix (see the THEORY section). The computation of the Amari index requires that the theoretical **A** matrix be known, which is not normally the case. As to interpretation, a small value of the Amari index is indicative of a satisfactory model. Indeed, the best possible model corresponds to $\hat{\mathbf{A}} = \mathbf{A}$, in which case P_{err} would be equal to 0. However, in reality, this ideal situation is not attained, and Monakhova et al. [35] indicate that a P_{err} value below 0.05 is obtained when a good decomposition is reached, while a P_{err} value above 0.2 corresponds to an "unacceptably poor performance".

Another method, called SONIC [36], standing for Simulated Ordered Negentropy of Independent Components, was recently proposed for the determination of the correct number of ICs. The SONIC method is based on the Gap statistic [37], which was originally developed to estimate the number of clusters in a multivariate data set. SONIC works by comparing the negentropy value of estimated ICs to its expected value (in so far as it is known). The negentropy is calculated as the difference between the entropy of a Gaussian variable and that of the measured variable, both variables having the same mean and standard deviation. It is always positive.

Although several methods have already been proposed to determine the optimal number of ICs, not all of them can be applied in every case: For example, as was mentioned earlier, the Amari index requires the mixing matrix (i.e., the proportions in which the pure components are mixed) be known, which is often not the case. The same is the case for the SONIC method, where the estimated negentropy value is compared to the expected one, implying that the negentropy is known, which again is not usually the case. Therefore, and as was pointed out recently by Poncela [38], "a further direction for research in ICA could be to develop a test statistic for the number of ICs." The objective of this article is to propose two fast and simple methods to determine the optimal dimensionality of ICA models, which do not require any knowledge of the mixing matrix **A** or of the pure signals, but which rely solely on the characteristics of the experimental data and the extracted vectors. These methods have been tested on different (simulated and real) data sets, and were compared to the method based on the computation of the SSR statistic, and to a slightly modified version of the method proposed by Westad et al.

2. Theory

2.1. Independent Components Analysis (ICA)

In Independent Components Analysis, one assumes the $m \times p$ **X** matrix can be decomposed as

$$\mathbf{X} = \mathbf{AS} \quad (4)$$

where **S** is the $k \times p$ matrix of k independent source signals (in rows), called the Independent Components, and **A** is the $m \times k$ mixing matrix. ICA aims at determining both **A** and **S**, knowing only **X**, by assuming that the source signals are mutually statistically independent, and that their mixing (yielding **X**) is linear. ICA, which aims to maximise independence between the extracted components, assumes that the "pure source signals" are less Gaussian than their mixtures (Central Limit Theorem) and so maximises the non-Gaussianity of the extracted ICs.

By analogy with other latent-variables based methods, such as PCA, the signals in the rows of **S** (the Independent Components) can be assimilated to loadings vectors, while the values in the rows of the mixing matrix **A,** which correspond to the proportions of the different pure signals in the mixed signal in each row, can be assimilated to scores vectors.

For this study, ICA calculations were done using the JADE algorithm [21].

2.2. Validation methods

2.2.1. The "ICA-by-Blocks" method

The method presented here starts by splitting the data matrix into B blocks of samples of approximately equal size (equal numbers of rows). Care must be taken in the construction of these blocks, so that the samples in each block are representative of the whole data matrix. The size of the blocks must be large enough to enable the computation of an ICA model with sufficient ICs (the maximum number of computed ICs, A_{max}, should exceed the expected optimal number of ICs).

For each of these predefined blocks, A_{max} ICA models are computed, with from 1 to A_{max} ICs. Since the signs of ICs of different models representing the same source signal may change from one model to another and from one block to another, the signs of the vectors in **A** and in **S** are adjusted so that the most intense value in each vector of **S** is positive. ICs corresponding to true source signals should be found in all representative subsets of samples, or row blocks, of the full data matrix. Such true ICs calculated from different blocks should be strongly correlated. When too many ICs are extracted, they will tend to contain noise characteristic of only that particular block. These noisy ICs will then have lower correlations with the ICs extracted from other blocks. Similarly, when too few ICs are extracted, they might correspond to mixtures of pure source signals in different proportions for different blocks, in which case the correlations between ICs from different blocks would be lower.

This procedure results in B models with 1 IC, B models with 2 ICs, and so on up to B models with A_{max} ICs. All the models computed with the same number, n, of ICs, ($1 \leq n \leq A_{max}$), are then compared by calculating the absolute correlation between each pair of pure signals, or ICA-loadings, forming the rows of the S matrices for the B different blocks. In the general case, for n-ICs ICA models, an $n.B \times n.B$ correlation

		B = 1			B = 2			B = 3		
		IC1	IC2	IC3	IC1	IC2	IC3	IC1	IC2	IC3
B = 1	IC1	1	0	0	↑				↑	
	IC2	0	1	0		↑		↑		
	IC3	0	0	1						
B = 2	IC1	↑			1	0	0		↑	
	IC2		↑		0	1	0	↑		
	IC3				0	0	1			
B = 3	IC1		↑			↑		1	0	0
	IC2	↑			↑			0	1	0
	IC3							0	0	1

Fig. 1. The correlation matrix of 3-ICs ICA models applied on the 3 Blocks of a data matrix where only 2 ICs are significant. The '↑' symbols represent large correlations.

matrix is obtained. The correlation of an IC with itself being equal to 1, the correlation matrix will have 1's along the diagonal. As well, the correlations between different ICs of the same block being theoretically equal to zero, the correlation matrix will also contain $(n^2-n)\times B$ values very close to zero. If all n ICs are significant, they will appear in each of the B models, although possibly in a different order, so that the correlation between equivalent ICs in different blocks will be close to 1. If too many ICs are extracted from the blocks, the extraneous ICs will contain a significant contribution related to noise, and so will be significantly less correlated to all of the ICs from the other blocks.

Each correlation matrix is then vectorised, and its elements are sorted in decreasing order to give what will be called a correlation-vector. The sorted correlation-vector is of length $(n\times B)^2$, with the first $n\times B$ values equal to 1. The correlation matrix being symmetrical, all the other values are duplicated.

As a simple example, let us assume that $B=3$, $n=3$ and the optimal number of ICs, $A_{opt}=2$. The correlation matrix can be represented as in Fig. 1.

The 9 $(n\times B)$ values on the diagonal correspond to the correlation of each IC with itself, and are therefore equal to 1.

As the optimal number of ICs, A_{opt}, is equal to 2, only two ICs of Block 1 are correlated to two ICs of Block 2 and to two ICs of Block 3. Therefore, in each off-diagonal block of the correlation matrix, a maximum of 2 values are "large" (close to 1), while all other values are negligible. As there are $(B^2 - B)$ off-diagonal blocs, there is a theoretical maximum of $n\times(B^2-B)$ (i.e. $3\times(9-3)=18$) "large" correlation values in these off-diagonal blocks. But, if $A_{opt}=2$, there are in fact only $2\times(9-3)=12$ "large" off-diagonal correlation values.

Therefore, in the sorted correlation-vector, the first $n\times B$ values are equal to one, and the next $n\times(B^2-B)$ values, at most, may be close to 1; all other values are negligible. In order to find the optimal number of factors, one does not need to plot the whole sorted correlation vector, but only part of it, from $(n\times B)+1$ to $(n\times B)+n\times(B^2-B)$. This corresponds to $n\times(B^2-B)$ correlation coefficients for equivalent ICs in different blocks. Because of the duplicate values, only every second point of this vector need be plotted.

When plotting these $(n\times(B^2-B))/2$ elements of the correlation-vector, a certain number, N, of values will be found to be close to 1, while the other values will be smaller. These N values come from the high correlation values in the $(B^2-B)/2$ off-diagonal blocks of the correlation matrix. As only every second point of the correlation

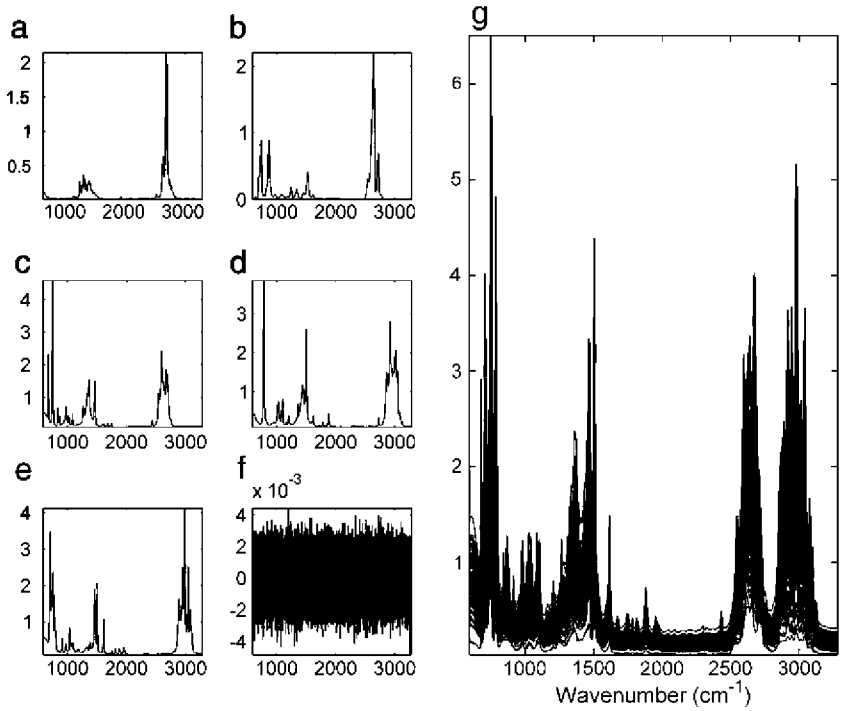

Fig. 2. Signal generated from the IR spectra of a) 2,6-octadiyne, b) 5-methylenebicyclo[2.2.1]-2 heptene, c) 1,3-dimethyl benzene, d) 1,4-dimethyle benzene, e) ethyl-benzene; and f) the random Gaussian noise matrix. (On these 6 subplots, the x axis is the approximate wavenumber range (in cm^{-1})); g) the Simulated data matrix.

vector is plotted, these values correspond to either the blocks of the upper-off-diagonal of the correlation matrix, or to the lower-off-diagonal. Each of these off-diagonal blocks contains $N/((B^2-B)/2)$ high correlation values. Therefore, A_{opt} is equal to $N/((B^2-B)/2)$.

It has already been pointed out that the ICs may not be extracted in the same order for different blocks. This means that a particular signal maybe extracted from one block and a different signal from another block. This would result in a decrease in the calculated correlation values. However, as the number of ICs extracted increases, both signals will be extracted for both blocks, and so the correlation values will increase again. Therefore, the optimal model will be defined as the model where all the correlations are relatively high.

2.2.2. Method based on the Durbin–Watson criterion

The method based on the Durbin–Watson (DW) criterion has recently been successfully applied to ICA models [39]. This criterion has been proposed as a measure of the signal/noise ratio in signals [40]. The value of the DW criterion applied to a signal **s** of length n is defined as:

$$\mathrm{DW} = \frac{\sum_{i=2}^{n} (s(i)-s(i-1))^2}{\sum_{i=1}^{n} s(i)^2} \tag{5}$$

where $\mathbf{s}(i)$ is equal to the value of the ith data point in **s**.

The value of the Durbin–Watson criterion tends to 0 when there is no noise in the signal, and tends towards 2 if the signal contains only noise. Rutledge et al. have used this criterion for the validation of multivariate models [41–43], in order to estimate numerically whether the latent variables (loadings) or the regression coefficients of the closed form of a PLS model (the so-called b-coefficients) were significant or not, i.e., whether they contained more information (structure) than noise.

When applied to Independent Components based on structured pure source signals (infrared spectra for example), the same is to be expected: as soon as a relatively large amount of noise is included in the IC, the DW value should increase, thus indicating the number of significant ICs as this number minus 1.

The procedure followed here consists in calculating A_{max} ICA models, with respectively from 1 to A_{max} ICs. The results of each ICA model are used to calculate A_{max} residual matrices containing the signals and noise remaining after deflating with from 1 to A_{max} ICs. For each sample, the DW values are calculated for each of these residuals matrices. The evolution of the DW values as a function of the number of ICs extracted can reveal the number of source signals within the mixed signal of each sample. For interpretation purposes, DW colour plots can be drawn, representing the evolution of the DW criterion for the residuals for each sample as the number of ICs included in the ICA models increases.

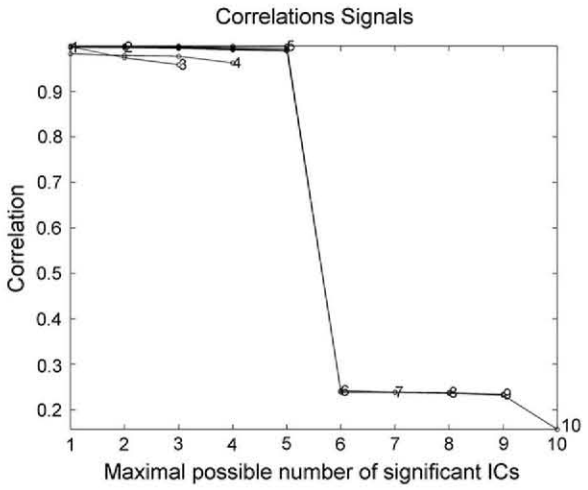

Fig. 3. Signal-Correlation graph for the Simulated data (2 Blocks, 1 to 10 ICs).

3. Experimental

3.1. The data

3.1.1. Simulated data

In order to simulate spectroscopic data, 5 InfraRed (IR) spectra (ethyl-benzene, 1,4-dimethyl benzene, 1,3-dimethyl benzene, 5-Methylenebicyclo[2.2.1]-2 heptene and 2,6-octadiyne) were downloaded from the NIST data base (National Institute of Standards and Technology) [44], transformed to have similar resolutions and approximate wavenumber ranges for the same total number of data points. These signals were then used to generate 100 spectral mixtures, their proportions being randomly generated. Randomly-generated Gaussian noise with a null mean and a 10^{-3} standard deviation was added to the resulting data matrix, corresponding to about 0.1% noise. The 5 source signals, the Gaussian noise matrix, as well as the resulting 100×800 signal matrix, are presented in Fig. 2.

This data set, for which we know the optimal number of ICs is 5, will be used to demonstrate the proposed methods.

3.1.2. Fluorescence of heated oils data

These data represent 3D Front-Face fluorescence spectra of corn oils, heated at different temperatures for different periods of time, and to which were added or not a natural (*Nigella sativa* L. seed extract), or a synthetic (butylated hydroxytoluene) antioxidant. The goal of this study was to evaluate the antioxidant effect of Nigella seed extract on the stability of edible oils during their accelerated thermal oxidation [39]. 320 spectra were available, measured at 111 excitation wavelengths (between 280 and 500 nm, with a step of 2 nm) and 126 emission wavelengths (between 300 and 550 nm, with a step of 2 nm). In order to perform the ICA calculations, the cubic data array was unfolded to give a $320\times13{,}986$ spectral data matrix.

3.1.3. Apple data

This data set has already been presented in details in a previous article [45]. It consists of 94 visible-IR spectra of apples, described by three characteristics, namely the cultivar (*Cox* or *Jonagold*), the

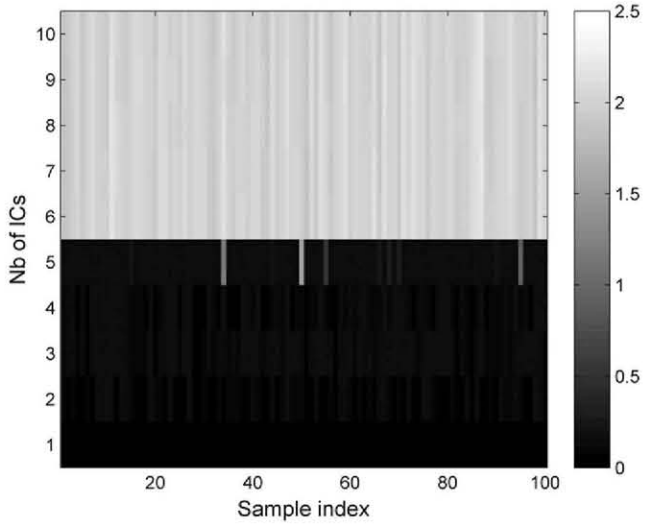

Fig. 4. Simulated data, the DW colour plot: Durbin–Watson values for the residues after subtracting signals calculated using ICA models with increasing number of ICs.

maturity level (Fresh, Medium or Mealy), and the colour of the face measured (Red, Green). Each reflection spectrum was the average of 5 individual optical scans from 380 to 2000 nm with 7.5 nm increments. SNV transformation was applied to the 94 spectra before analysis. Both the red side and the green side of the apples were measured. Therefore, one can expect three levels of clustering in the data, depending on the cultivar, the maturity level, and the face measured.

3.2. Software

All computations were performed using Matlab 7.6.0 (R2008a) (The MathWorks Inc., Natick (MA, USA), 2008), with the JADE algorithm downloaded from ref. [46], and in-house codes for the validation methods.

4. Results and discussion

4.1. Simulated data

The ICA-by-Blocks method was applied on these data, by splitting the data into 2 blocks in a venetian-blind fashion (1, 2; ...; 1, 2), with from 1 to 10 ICs. The "signal-correlation" plot is presented in Fig. 3.

The number at the extremity of each curve indicates the number of ICs calculated in the model considered. As can be seen, up to the 5-ICs model, the correlations between corresponding ICs from each block are close to 1, indicating that similar ICs are extracted in each segment. Therefore, one concludes that these ICs are significant. Addition of more ICs to the model leads to very poor correlations between ICs of the two blocks, indicating that the extracted ICs are not valid signals. Therefore, one can conclude that the optimal number of ICs in this data set is equal to 5. These results are in agreement with what was expected.

4.1.1. Influence of the number of blocks

When applying this method, the user has to input two parameters, namely the maximal number of ICs, A_{max}, and the number of blocks, B. In order to choose B, intuitively, one could say that each block should be as representative of the whole sample population as possible, implying it should contain as many samples as possible, so that two blocks would be an optimal number. In order to assess this assumption, the ICA-by-Blocks method was applied several times to the simulated data, by setting the maximum number of ICs to 10, and by varying the number of Blocks ($B=4$, 6, 10 were tested, in addition to $B=2$ presented above). Each time, a figure similar to the one presented above was obtained, and so 5 ICs were found as the optimal number. Therefore, for the reasons explained earlier, the smallest value of 2 blocks will be used. As to choosing A_{max}, one can rely on prior knowledge of the data when available (physico-chemical knowledge, for example).

The method based on the Durbin–Watson criterion was also applied on the complete data set with from 1 to 10 ICs, and the resulting DW colour plot is presented in Fig. 4.

This figure represents the value of the DW for the residual signal of each of the 100 samples after removing the calculated contribution of each of the 10 ICA models. Except for a few samples for which a 4-ICs model would appear to be optimal, it is clear that the optimal number of ICs is 5.

4.1.2. Influence of the level of noise

In order to test the robustness of the method as a function of the noise level, and because the noise level was relatively low, the level of Gaussian noise added to the Simulated data was multiplied by 100 (Fig. 5a), yielding a level of noise of about 10%. The ICAbyBlocks and DW method were applied to this very noisy data set (Fig. 5b) and c), respectively).

Fig. 5b shows that the correlations between the ICs of each block are slightly lower than when the data is much less noisy, and indeed, the presence of noise in the ICs can be seen in Fig. 5c. However, both methods continue to indicate 5 ICs as the optimal number.

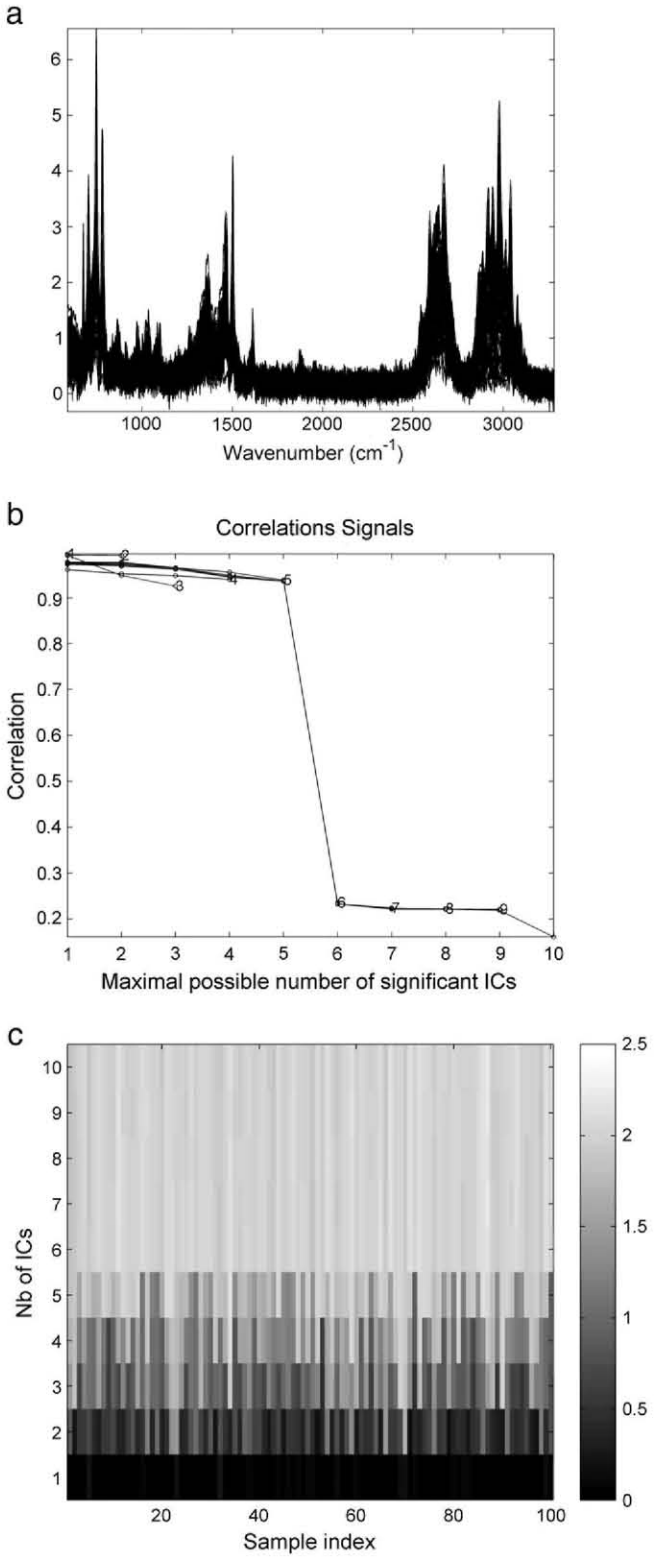

Fig. 5. a) The noisy simulated data; b) Signal-Correlation graph for the noisy Simulated data (2 blocks, 1 to 10 ICs); c) The DW colour plot for the noisy Simulated data.

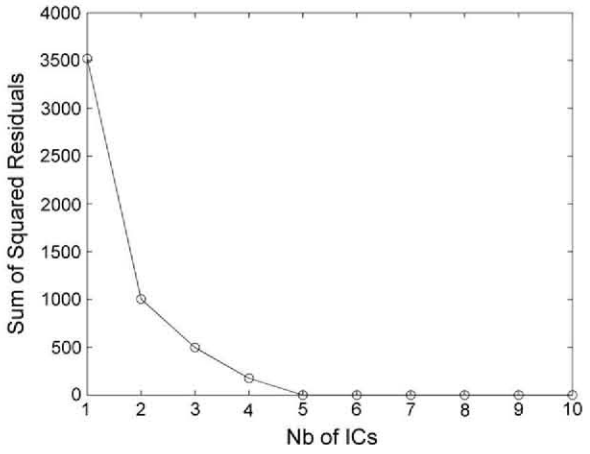

Fig. 6. Simulated data: Sum of Squared Residuals as a function of the number of ICs used to reconstruct the data matrix.

4.1.3. Method based on the Sum of Squared Residuals (SSR) of the reconstructed data matrix

10 ICA models were calculated, with from 1 to 10 ICs, and each time the original data matrix was reconstructed, and the residuals were calculated. The plot showing the SSR as a function of the number of ICs used to reconstruct the data matrix is presented in Fig. 6.

Here again, the results are as expected: the SSR decreases regularly until 5 ICs, after which it remains stable. (Similar results are obtained with the noisy simulated data, but with larger SSR values).

4.2. Fluorescence of heated oils data

The ICA-by-Blocks method was applied to this data set, with $B = 2$ blocks and $A_{max} = 20$. The correlation plot is presented in Fig. 7a.

It can be seen that after extracting 7 ICs, the curves go down progressively through 8, 9 and 10, but then start going up again to 17 ICs. By adding more than 7 ICs, up to 10 ICs-models, the correlations between the ICs of the different blocks are much lower, indicating either that noise is being extracted, so that 7 ICs is optimal, or that different significant ICs are extracted from each block, leading to relatively low correlations between them. As more than 10 ICs are introduced in each model, more ICs extracted in each block are similar to the ICs extracted in the other block, leading to increasing correlation values, up to 17 ICs, which is thus the optimal number of ICs in this data set.

This evolution can be understood in the light of the results in Fig. 7b, which shows the Durbin–Watson (DW) values for the residues after subtracting the signals calculated using ICA models with increasing numbers of ICs. White corresponds to noisy signal residues, black to smooth signal residues. The graphic corresponds to the average of the

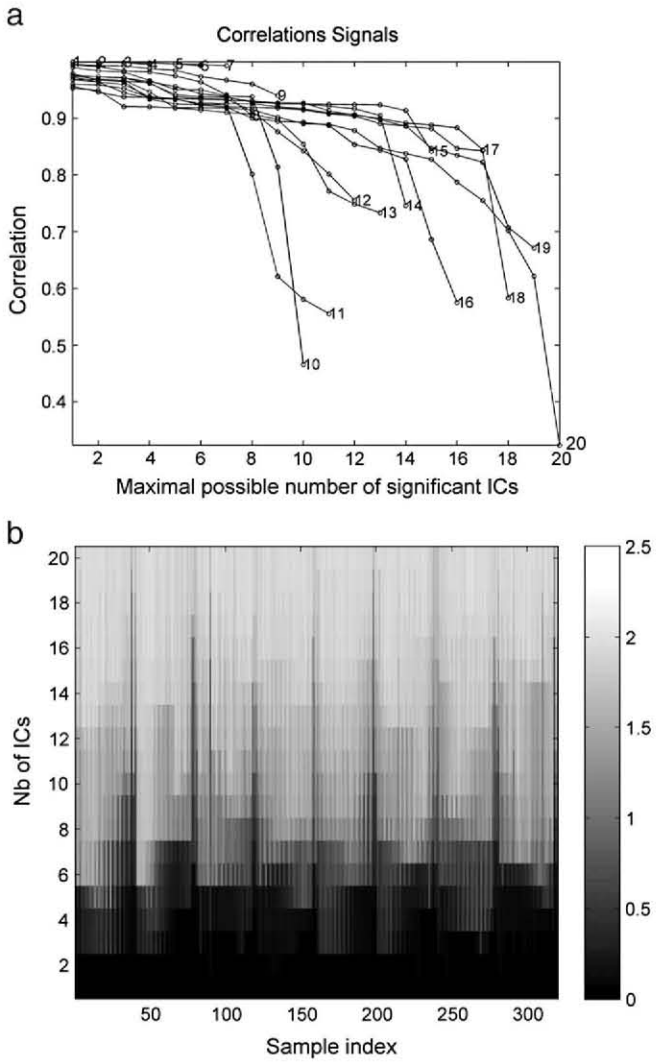

Fig. 7. Fluorescence data: a) ICA-by-Blocks ($B = 2$ and $Amax = 20$); b) DW colour plot.

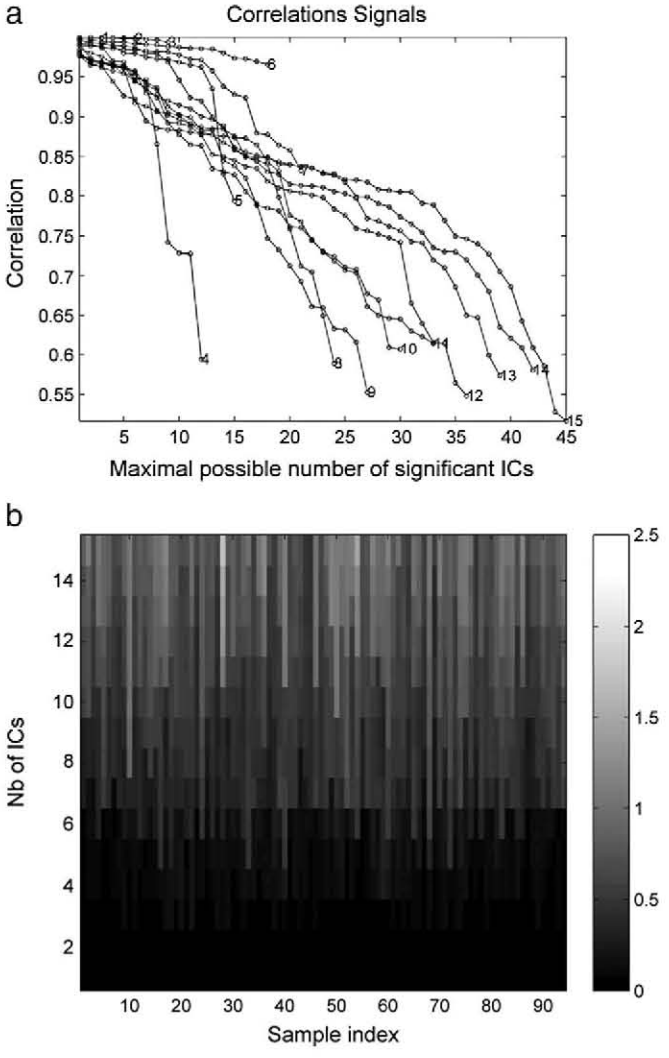

Fig. 8. Apple data: a) ICA-by-Blocks with 3 blocks (Venetian-blind selection) and 15 computed ICs; b) the DW colour plot.

DW values for the residues of the unfolded spectrum of each of the 320 samples. It can be seen that most samples only contain noise after subtracting more than 10 ICs. Some samples however, are still structured after subtracting 15 ICs. In ref. [39], the colour maps of the DW values averaged over all 320 samples, for each excitation and emission wavelength, respectively, were presented. It could be seen that a small number of wavelengths required more than 10 ICs to extract all the structured signals from the data. In a situation such as this, it is to be expected that blocks will extract minor source signals in a different order.

The method based on SSR was also applied to this data set, and the optimal number of ICs was found to be 7. However, Ammari et al. [39] have shown that at least 16 ICs were significant, and carried important information concerning the complex chemical reactions occurring during the heating of the oil. Therefore, in this case, the SSR method fails to find an appropriate optimal number of ICs.

4.3. Apple data

The ICA-by-Blocks method (3 blocks with venetian-blind selection) and the DW criterion were calculated on the apple data, and the results are presented in Fig. 8a and b. It was necessary to use 3 blocks here to have comparable blocks, as the samples were sorted in alternating order for face colour.

With ICA-by-Blocks, the models built with up to 3 ICs in each block lead to very correlated ICs in the three blocks. When adding a fourth and a fifth IC to the model, the ICs calculated in the three blocks are clearly not correlated. Addition of a 6th IC to the model increases the correlation between the ICs of the three blocks to more than 0.95. Adding more ICs to the model worsens the correlations. Therefore, one could hesitate between a 3-ICs model and a 6-ICs model. This is confirmed by the DW method: The DW colour plot shows that with 3 ICs, the DW criterion is very low. It increases slowly up to 6 ICs, but its value remains relatively low, and increases more rapidly afterwards.

As the correlation graph shows a very large correlation for the 6-ICs model, 6 ICs were kept as optimal. The 6-ICs ICA model was calculated, and the IC-loadings and scores are presented in Fig. 9.

Investigation of the loading plots indicate that IC4 to IC6 clearly accounts for pigments, IC4 corresponding to chlorophyll. IC1 to IC3 could be related to structural changes in the apples as they evolve from non-mealy to mealy [47]. As to the scores, although ICs 2, 3 and 6 all separate to a certain extent the apples based on the colour of the face measured, the best separation is observed for IC4. IC5 separates the samples according to the maturity level (the Fresh and Mealy levels are well separated, while the intermediate Medium level slightly overlaps on both). As to the type of apple, IC2 and IC4 both contain information about the variety and these two ICs clearly separate the *Jonagold* samples from the *Cox* samples.

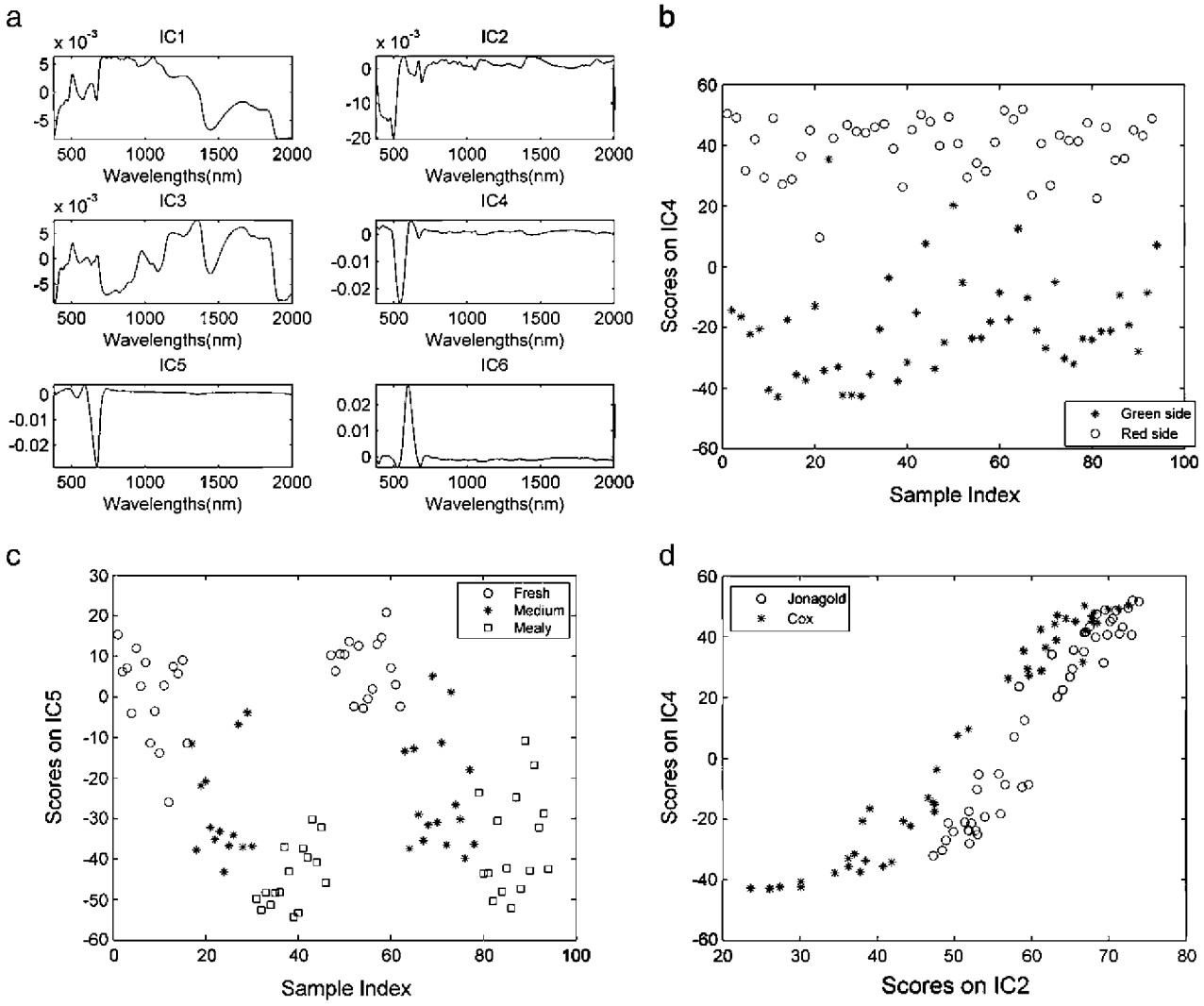

Fig. 9. Apple data: a) The first 6 ICA-loadings; b) ICA-scores on IC4 *vs* sample index; c) ICA-scores on IC5 *vs* sample index; d) IC2 *vs* IC4 score plot.

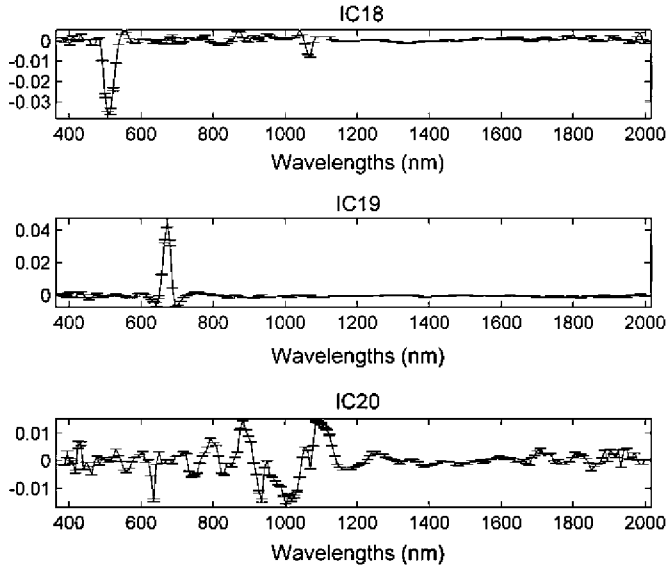

Fig. 10. Coefficients and Westad's uncertainty parameter for each variable in ICs 18 to 20 calculated using the modified Westad method on the Apple data, with 3 blocks (Venetian-blind selection) and 20 computed ICs.

4.3.1. Comparison with the modified Westad Method [18]

The ICA-by-Blocks and DW methods were compared to the method proposed by Westad on the Apple data. However, this latter method had to be slightly modified beforehand.

The method proposed by Westad is a cross-validation of ICA, with the calculation of an uncertainty parameter to help decide on the number of ICs to use in the final model. Cross-validation is often used in validation of multivariate models, but is known to sometimes lead to overfitted models, and to under-estimate the actual prediction error related to the final model: indeed, if S cross-validation segments are used, each sample contributes to the computation of S-1 sub-models, so that not only significant variation, but also part of the noise, is modelled. When enough samples are available, independent test set validation is to be preferred. In this work, in order to be consistent with what is done in the ICA-by-Blocks method, and to get a fair comparison, cross-validation will be replaced by splitting the data set into the same few blocks for the calculation of the uncertainty parameter. The results obtained should be less optimistic than those with cross-validation.

Another aspect of the method proposed by Westad is the use of the Procrustes rotation to solve the problem of ICs possibly being extracted in different orders and signs when different cross-validation segments are removed. However, the user cannot be sure that a particular IC is extracted from one CV-block, but not from another. When the sets of ICs obtained after the deletion of two cross-validation segments are different, Procrustes rotation cannot work properly. To avoid this problem and still have the ICs in the same order and with the same signs, in this work, the ICs computed in each block are first oriented so that similar ICs in different blocks have the same sign, and are then sorted according to their correlation with the ICs extracted from the original data matrix, so that even if some ICs are not extracted in all blocks, they are not modified.

This slightly modified Westad method was applied to the apple data (with a maximum of 20 computed ICs), with 3 blocks and a Venetian-blind selection. Only the three last ICs of the 20-ICs model are presented in Fig. 10.

Clearly, ICs 18 to 20 seem to have very significant loading values, as the confidence interval around each data point (Westad's uncertainty parameter) is very narrow. This is the case for all 20 extracted ICs. However, one expects the dimensionality of the ICA model to be much lower as there are only 3 controlled factors, possibly with interactions. Both the ICA-by-Blocks and the DW methods found 6 ICs. Similar results were had with other data sets (not shown here). Therefore, we believe this method overfits, and is therefore less reliable than ICA-by-Blocks or Durbin–Watson.

5. Conclusion

Two novel methods were proposed to determine the optimal number of ICs to use in an ICA model. These methods are simple, relatively fast, and do not require any specific prior knowledge about the data, which make them methods of choice in many cases. The criterion based on the correlations between blocks can be applied to any type of data. For this method, care must be taken to generate comparable, representative data blocks. The Durbin–Watson signal/noise criterion can only be applied to structured signals.

The results obtained on simulated data and in 2 real case studies are in agreement with what was expected. Investigation of the scores and loadings vectors obtained for the final models were shown to be relevant.

References

[1] A. Hyvärinen, E. Oja, Independent component analysis: algorithms and applications, Neural Networks 13 (4–5) (2000) 411–430.
[2] L. De Lathauwer, B. De Moor, J. Vandewalle, An introduction to independent component analysis, Journal of Chemometrics 14 (2000) 123–149.
[3] G. Wang, Q. Ding, Z. Hou, Independent component analysis and its applications in signal processing for analytical chemistry, TrAC 27 (4) (2008) 368–376.
[4] S.P. Huang, C.C. Chiu, Process monitoring with ICA-based signal extraction technique and CART approach, Quality and Reliability Engineering International 25 (2009) 631–643.
[5] G.L. Wallstrom, R.E. Kass, A. Miller, J.F. Cohn, N.A. Fox, Automatic correction of ocular artifacts in the EEG: a comparison of regression-based and component-based methods, International Journal of Psychophysiology 53 (2004) 105–119.
[6] V. Krishnaveni, S. Jayaraman, P.M. Manoj Kumar, K. Shivakumar, K. Ramadoss, Comparison of independent component analysis algorithms for removal of ocular artefacts from electroencephalogram, Measurement Science Review 5 (2) (2005) 67–78.
[7] F. Szabo de Edelenyi, A.W. Simonetti, G. Postma, R. Huo, L.M.C. Buydens, Application of independent component analysis to ^{1}H MR spectroscopic imaging exams of brain tumours, Analytica Chimica Acta 544 (2005) 36–46.
[8] J. Hao, X. Zou, M.P. Wilson, N.P. Davies, Y. Sun, A.C. Peet, T.N. Arvanitis, A comparative study of feature extraction and blind source separation of independent component analysis (ICA) on childhood brain tumour ^{1}H magnetic resonance spectra, NMR in Biomedicine 22 (2009) 809–818.
[9] S. Umeyama, Blind deconvolution of blurred images by use of ICA, Electronics and Communications in Japan Part 3 84 (12) (2001) 1–9.
[10] D. Jouan-Rimbaud Bouveresse, H. Benabid, D.N. Rutledge, Independent component analysis as a pretreatment method for parallel factor analysis to eliminate artefacts from multiway data, Analytica Chimica Acta 589 (2007) 216–224.
[11] C. Di Natale, E. Martinelli, A. D'Amico, Counteraction in environmental disturbances of electronic nose data by independent component analysis, Sensors and Actuators B: Chemical 82 (2002) 158–165.
[12] F.C. Meinecke, S. Harmeling, K.R. Müller, Inlier-based ICA with an application to superimposed images, International Journal of Imaging Systems and Technology 15 (1) (2005) 48–55.
[13] M. Vosough, Using mean field approach independent component analysis to fatty acid characterization with overlapped GC-MS signals, Analytica Chimica Acta 598 (2007) 219–226.
[14] G. Wang, Y.A. Sun, Q. Ding, C. Dong, D. Fu, C. Li, Estimation of source spectra profiles and simultaneous determination of polycomponent in mixtures from ultraviolet spectra data using kernel independent component analysis and support vector regression, Analytica Chimica Acta 594 (2007) 101–106.
[15] F. Liu, Y. Jiang, Y. He, Variable selection in visible/near infrared spectra for linear and nonlinear calibrations: a case study to determine soluble solids content of beer, Analytica Chimica Acta 635 (2009) 45–52.
[16] N. Pasadakis, A.A. Kardamakis, Identifying constituents in commercial gasoline using Fourier transform-infrared spectroscopy and independent component analysis, Analytica Chimica Acta 278 (2006) 250–255.
[17] A.A. Kardamakis, A. Mouchtaris, N. Pasadakis, Linear predictive spectral coding and independent component analysis in identifying gasoline constituents using infrared spectroscopy, Chemometrics and Intelligent Laboratory Systems 89 (2007) 51–58.
[18] F. Westad, M. Kermit, Cross validation and uncertainty estimates in independent component analysis, Analytica Chimica Acta 490 (2003) 341–354.
[19] S. Wold, K. Esbensen, P. Geladi, Principal Component Analysis, Chemometrics and Intelligent Laboratory Systems 2 (1987) 37–52.
[20] A. Hyvärinen, E. Oja, A fast fixed-point algorithm for ICA, Neural Computation 9 (1997) 1483–1492.

[21] J.F. Cardoso, A. Souloumiac, Blind beamforming for non-Gaussian Signals, IEE Proceedings F: Radar and Signal Processing 140 (6) (1993) 362–370.
[22] A.J. Bell, T.J. Sejnowski, An information maximisation approach to blind separation and blind deconvolution, Neural Computation 7 (1995) 1129–1159.
[23] P.A.D.F.R. Højen-Sørensen, O. Winther, L.K. Hansen, Mean-Field approaches to independent component analysis, Neural Computation 14 (2002) 889–918.
[24] R.B. Francis, I.J. Michalel, Kernel independent component analysis, Journal of Machine Learning Research 3 (2003) 1–48.
[25] J. Yang, X.M. Gao, Y.J. Yang, Kernel ICA: an alternative formulation and its application to face recognition, Pattern Recognition 38 (10) (2005) 1784–1787.
[26] H. Stögbauer, A. Kraskov, S.A. Astakhov, P. Grassberger, Least-dependent-component analysis based on mutual information, Physical Review E: Statistical, Nonlinear, and Soft Matter Physics 70 (2004) 066123.
[27] http://www.klab.caltech.edu/~kraskov/MILCA/.
[28] S.A. Astakhov, H. Stögbauer, A. Kraskov, P. Grassberger, Monte Carlo algorithm for least dependent non-negative mixture decomposition, Analytical Chemistry 78 (5) (2006) 1620–1627.
[29] http://www.cs.umass.edu/~elm/ICA/.
[30] T. Naes, H. Martens, Principal component regression in NIR analysis: viewpoints, background details and selection of components, Journal of Chemometrics 2 (1988) 155–167.
[31] P. Geladi, B.R. Kowalski, Partial least-squares regression: a tutorial, Analytica Chimica Acta 185 (1986) 1–17.
[32] M. Stone, Cross-validatory choice and assessment of statistical predictions, Journal of the Royal Statistical Society B36–2 (1974) 111–133.
[33] S. Wold, Cross-validatory estimation of the number of components in factor and principal components models, Technometrics 20 (4) (1978) 397–405.
[34] G. Wang, W. Cai, X. Shao, A primary study on resolution of overlapping GC-MS signal using mean-field approach independent component analysis, Chemometrics and Intelligent Laboratory Systems 82 (2006) 137–144.
[35] Y.B. Monakhova, S.A. Astakhov, A. Kraskov, S.P. Mushtakova, Independent components in spectroscopic analysis of complex mixtures, Chemometrics and Intelligent Laboratory Systems 103 (2010) 108–115.
[36] S.M. Lee, Estimating the number of Independent Components via the SONIC Statistic, Master of Science in Applied Statistics Thesis Dissertation, Oxford (UK) 2003.
[37] R. Tibshirani, G. Walther, T. Hastie, Estimating the number of clusters in a data set via the gap statistic, Journal of the Royal Statistical Society 63 (2) (2001) 411–423.
[38] P. Poncela, Further research on independent component analysis, International Journal of Forecasting 28 (1) (2012) 94–96, doi:10.1016/j.ijforecast.2011.02.004.
[39] F. Ammari, C.B.Y. Cordella, N. Boughanmi, D.N. Rutledge, Independent Component Analysis applied to 3D-Front Face Fluorescence of spectra of edible oils to study the antioxidant effect of *Nigella sativa* L. extract on the thermal stability of heated oils, Chemometrics and Intelligent Laboratory Systems (2011), doi:10.1016/j.chemolab.2011.06.005.
[40] J. Durbin, G.S. Watson, Testing for serial correlation in least squares regression, Biometrika 37 (1950) 409–428.
[41] D.N. Rutledge, A.S. Barros, Durbin–Watson statistic as a morphological estimator of information content, Analytica Chimica Acta 454 (2002) 277–295.
[42] S. Gourvénec, D.L. Massart, D.N. Rutledge, Determination of the number of components during mixture analysis using the Durbin–Watson criterion in the Orthogonal Projection Approach and in the SIMPLe-to-used Interactive Self-modelling Mixture Analysis approach, Chemometrics and Intelligent Laboratory Systems 61 (2002) 51–61.
[43] M.P. Gomez-Carracedo, J.M. Andrade, D.N. Rutledge, N.M. Faber, Selecting the optimal number of partial least squares components for the calibration of attenuated total reflectance mid-infrared spectra of undersigned kerosene samples, Analytica Chimica Acta 585 (2007) 253–265.
[44] http://webbook.nist.gov/chemistry/form-ser.html.
[45] D. Jouan-Rimbaud Bouveresse, A.S. Barros, D.N. Rutledge, Generalised PLS_Cluster: an extension of PLS_Cluster for interpretable hierarchical clustering of multivariate data, Sensing and Instrumentation for Food Quality and Safety 1 (3) (2007) 79–90.
[46] http://perso.telecom-paristech.fr/~cardoso/Algo/Jade/jadeR.m.
[47] P. Barreiro, A. Moya, E. Correa, M. Ruiz-Altisent, M. Fernández-Valle, A. Peirs, K.M. Wright, B.P. Hills, Prospects for the rapid detection of mealiness in apples by non-destructive NMR relaxometry, Applied Magnetic Resonance 22 (2002) 387–400.