10th International IFAC Symposium on
Computer Applications in Biotechnology
Preprints Vol.1, June 4-6, 2007, Cancún, Mexico

UNRAVELLING SPECTRAL SIGNATURES IN BIOPROCESS DEVELOPMENT

Hacer Kilic, Elaine Martin and Gary Montague

School of Chemical Engineering and Advanced Materials, Newcastle University, Newcastle upon Tyne, NE1 7RU, England

Abstract. Fresh challenges are now arising for the bio-industries as a consequence of advances in Process Analytical Technologies (PAT) and the resulting generation of high dimensional spectral data. In particular, there is a need for analysis algorithms that realise information that will provide new and enhanced insights into bioprocess development. In this paper, a methodology is proposed which involves the application of Independent Component Analysis (ICA) to spectral signatures, following the appropriate pre-processing of the signals. An evolving window approach is considered to identify the number of key components present and serves to indicate critical operational points, i.e. their limitation or appearance within a batch. Also of interest is the use of the technique as a finger printing tool to help identify differences between batches as they evolve. The critical challenge is to make use of such methods in early stage development where typically only limited data is available. To place the technique proposed in context, the methodology is applied to NIR spectra generated from a pilot scale industrial antibiotic fermentation.

Keywords: Near Infrared Spectroscopy, Fermentation, Independent Component Analysis, Batch Monitoring

1. INTRODUCTION

In the bioprocess development environment, understanding the predominant reactions that are taking place at various stages throughout the course of a batch is a key knowledge objective. As changes in operating policy are explored, it is necessary to determine when important reaction pathways are significant and when they are impacted on by nutrient limitations or excessive accumulation. From such knowledge, operational policy changes can be made and new avenues of operation explored. The approach traditionally adopted involves the use of off-line sample analysis to identify the concentrations of nutrients that are perceived to be influential by the process scientists. Off-line sample analysis has its limitations, particularly with regard to the limited frequency at which data is available. Furthermore, the predominant reactions, and the reactants that drive them, can change significantly over the duration of a batch. A move towards more innovative solutions, to further the understanding of the chemical process, is of crucial importance if more rapid progression through the development cycle is to be achieved.

Bioprocess systems are typified by complex reaction mechanisms where, as of necessity, over-simplified kinetic descriptions are hypothesised. Even then bioprocess understanding of such mechanisms has traditionally been limited by the lack of informative on-line measurement. This situation has started to change as a result of the FDA initiatives in Process Analytical Technology (PAT). More specifically, the implementation of spectroscopic instrumentation alongside the application of advanced chemometric tools provides a real opportunity for the extraction of information that is both accepted by the regulatory authorities and which can be acted on. Furthermore it provides a route to deliver enhanced levels of bioprocess understanding.

Researchers have considered how PAT can be implemented to provide an on-line indication of products and metabolites through the construction of multivariate calibration models (Vaidyanathan,

2001). Although such information can be extremely useful, the focus of this paper differs in that it aims to identify the appearance or limitation of critical components. Based on knowledge of their spectral signatures, it can be ascertained how the influential components vary and hence how reaction schemes change throughout the progression of a batch. The methodology applied involves the initial application of Independent Component Analysis (ICA) to the spectral signatures, following appropriate pre-processing of the signals. In this paper an evolving window approach is considered.

A limitation of ICA is that the ordering of independent components, in terms of their extraction from the data, is not unique. This is in contrast to the multivariate statistical projection based technique of principal component analysis (PCA). In PCA the first principal component explains the greatest amount of variability in the data set with the next principal component defining the next greatest amount of variability and so on. Consequently there is natural ordering of the components. In ICA, the ordering of the ICs can change for each application of the algorithm as no constraints are imposed that ensure natural ordering. Consequently to understand the changes in behaviour over time, an analysis of the independent components for each window requires to be undertaken that recognises and addresses this limitation of ICA.

2. INDEPENDENT COMPONENT ANALYSIS

From the analysis of the signals obtained in neural activity assessment (Vigario et al, 1998; Ladroue et al, 2002) to the interpretation of financial information (Back and Weigend, 1997), there is a need to understand the underlying fundamental signals and the deviations that materialise in system changes. One technique that addresses these challenges that has recently been receiving significant interest in the literature is that of Independent Component Analysis (ICA). ICA was first proposed by Comon (1994). More recently Hyvärinen et al (2001) provided a comprehensive description of the theoretical background to ICA.

In the bio-industries, the application of ICA to provide an indication of the presence of key analytes from signals generated from complex measurement devices has been considered by a number of researchers. For example, Sholtz et al (2004) used ICA to fingerprint biological samples from a plant test system using spectral data from microchip-based nanoflow-direct-infusion QTOF mass spectrometry. The application of ICA was motivated by its ability to handle small data sets and it was shown to outperform principal component analysis (PCA). Lee and Batzoglou (2003) also showed that ICA outperformed PCA when applied to the analysis of clusters obtained from micro-array data generated from two example micro-organisms.

In this paper, the application of ICA to NIR spectral data is proposed. The hypothesis is that through the application of ICA, a linear transformation of the original spectral data will reveal the individual spectral signature of the pure compounds present within the process.

In terms of the basic ICA algorithm, the observed spectral data is assumed to arise as a consequence of a weighted linear combination of the pure individual species, the so called independent components, that are denoted by **s**. The weightings are termed the mixing coefficients (**A**). Thus the observed spectra, **x**, is denoted by:

$$\mathbf{x} = \mathbf{A}\mathbf{s} \tag{1}$$

Using the observed variables, i.e. the spectra in this paper, there is a requirement to estimate both the independent components that refer to the pure components together with their corresponding mixing coefficients. The ability to estimate **s** without prior knowledge of **A** requires a number of assumptions to be made. The two key assumptions are that the independent components are statistically independent and that they have non-Gaussian distributions which need not be known. The problem necessitates the determination of $\hat{\mathbf{s}}$:

$$\hat{\mathbf{s}} = \mathbf{W}\mathbf{x} \tag{2}$$

where **W** is the separating matrix which is a linear transformation of **x**. However since it is assumed that the independent components are uncorrelated, **W** can be calculated such that the mutual information of **s** is minimised. The problem can also be defined in terms of entropy, that is, from an information theoretic perspective. This approach results in the same solution strategy. The ICA algorithm involves a numeric search for a separating matrix which satisfies the information theoretic objectives. Since the search for **W** is typically initialised from a random starting position, the resulting solution is not unique. Furthermore, in the search for **W**, the sign is not relevant as it is the magnitude that is important. The search algorithms used in ICA are mainly divided into two types, gradient and fixed point. A typical example of the latter is the *FastICA* algorithm (Hyvärinen 1999). FastICA is used in the study reported in the paper.

3. EXPERIMENTAL DATA AND RESULTS

NIR spectra generated from a series of experimental design trials, undertaken on a pilot scale industrial antibiotic fermentation, formed the basis of the study into the applicability of ICA for spectral data

interpretation. The process used in the study is an industrial pilot-plant scale fermentation involving two stages, a seed stage and a production stage. Biomass is grown in the seed stage before being transferred to the final stage for the production of the desired product. The final stage is a fed batch process and lasts approximately 140 hours.

NIR measurements were collected on-line from the final stage of the process and formed the basis of the subsequent analysis. The spectral data were recorded every fifteen minutes using a Zeiss Corona 45 NIR. The instrument was operated in reflectance mode with the measurements lying in the range 950 – 1700 nm with a resolution of 6 nm. The instrument was equipped with a diode array detector of focal length 13 mm with a sampling area diameter of 15 mm and 15 detection fibres were situated around the inner edge of the lens. Multiple analyte concentrations, such as product, sugar, phosphate, lipids, ammonia, pH, viscosity and urea, were measured by off-line assay during the course of the fermentation. Data was available from seven batches in which natural variation resulted in a degree of variability. These batches potentially provide insight into behaviour that might be expected in a production environment where consistency of operation is sought. It is important in such operations to understand the causes of variation and act to reduce deviations. ICA offers the opportunity to gain insight into possible causes.

3.1 Spectral Data Pre-treatment

Fig. 1 shows the raw NIR spectra as gathered from the instrument after pre-treatment for an example batch. An outlier was removed using linear interpolation and the region 1650-1700nm was cut from the data set as it was at the extreme range of the instrument's detectability and was subsequently noisy. The multiple curves correspond to the NIR signals recorded throughout the evolution of the batch.

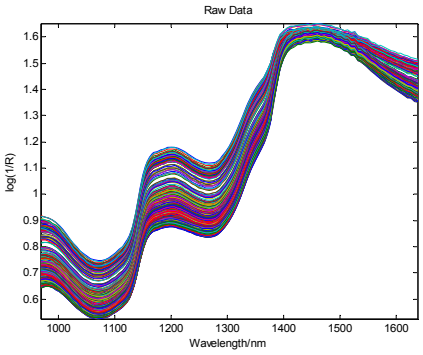

Fig. 1. Evolution of raw NIR spectra throughout a batch

First derivatives of the spectra (as shown in Fig. 2) were taken to remove the baseline drift that can occur with such instruments. The first derivatives were calculated using Savitsky-Golay smoothing (Gorry, 1990) for an 11 point window and a second order polynomial. These settings were found to be sufficient to achieve noise reduction whilst maintaining signal information content.

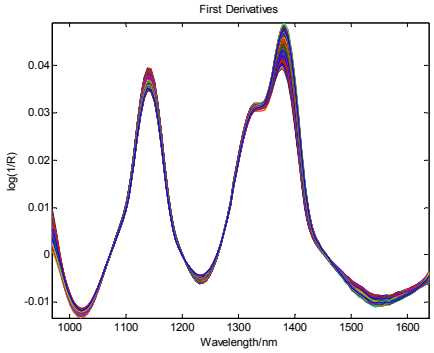

Fig. 2. First derivatives of NIR spectra

3.2 Independent Component Analysis Results

The next step in the analysis was to apply ICA to the pre-treated NIR data. The focus of the analysis was to ascertain the functionality of the ICA approach. Initially it was applied to verify whether it could extract the component spectra for compounds that are present throughout the course of the batch. It was applied to examine whether it is possible detect the appearance of a compound, in this case the product of interest. Finally by adopting a reverse approach, the algorithm was applied to see if it could detect the limitation of a compound. In this case, sugar was considered.

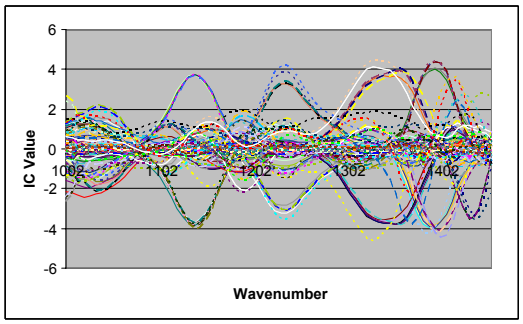

Fig. 3. All independent components for all windows for a single batch

An expanding window approach was utilised where the data presented to the ICA algorithm was the spectra from 0-10 hours, 0-20 hours until the window expanded to cover the whole batch, Fig. 3. Adopting this approach, in theory the spectra relating to those compounds present throughout the batch would be expected to be present in all windows, while the spectra relating to the product should appear later in the batch. The number of ICs retained for a particular window differed, likewise the number of IC's differed between batches for a particular window,

and was selected using both process understanding and through the examination of the IC's, i.e. only those IC's were retained where the information content was significant and no noise was captured. As discussed earlier, the functionality of the ICA algorithm is such that the ordering of the ICs is not consistent and the sign of the ICs is not significant. Consequently the results shown in Fig. 3 require further processing and interpretation to realise information extraction.

For this case study, it is known that biomass and lipids are present in significant concentrations throughout the course of the batch. Biomass can be predominantly detected at 1386nm and lipids at 1139nm. Figs. 4 and 5 show the ICs that have significant peaks in these two regions.

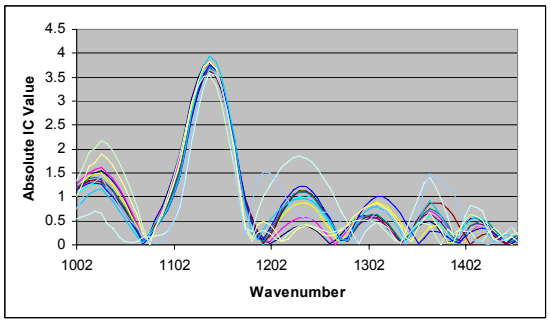

Fig. 4. ICs corresponding to lipids evolving over the course of a batch

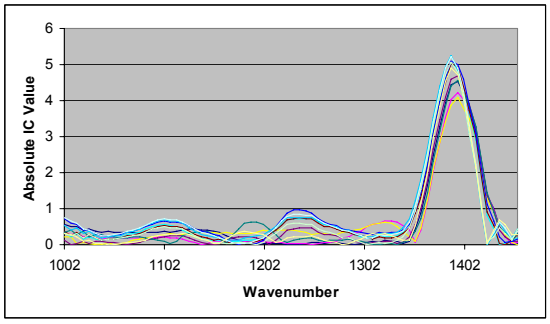

Fig. 5. ICs corresponding to biomass evolving over the course of a batch

Since the IC refers to the pure compound spectral signature, it would be expected that the magnitude of the IC would be consistent throughout the batch. However the results in Figs. 4 and 5 indicate variations do occur. At 1139 nm in Fig. 4 there is some variation in the magnitude of the IC signal and at 1386 nm in Fig. 5 there is somewhat greater variation in the biomass signal. This is to be expected as the changing conditions in the reactor such as biomass morphology impacts on the light scattering characteristics and therefore the IC signal obtained.

Figs. 6 and 7 show the magnitude of absorbance at 1386nm and 1139nm for all ICs, respectively. In Figs. 6 and 7, the dashed vertical lines separate the ICs for the expanding window (e.g. 0-10 hours, 0-20 hours etc). Around six ICs are obtained for each time window. It should be noted that as the sign of the IC is not significant, the absolute value has been reported. If the IC relating to product was present then it would be expected that one of the set of six ICs obtained for a specific time window would be of large magnitude, i.e. lie above the 95% confidence limit (horizontal line) whilst the remaining IC's for that window would be small. If the product wavenumber was not present then all ICs in the time window would be of low magnitude, i.e. below the 95% confidence limit. Fig. 6 shows the IC moving window results for lipids at 1139 nm and Fig. 7 the corresponding results for biomass at 1386 nm. In both cases it can be observed that a peaks lies above the confidence limit for each window for the duration of the batch.

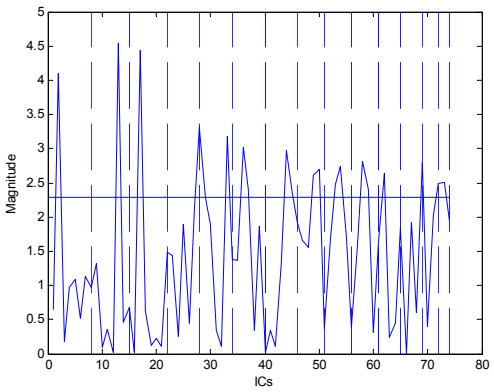

Fig. 6. Expanding window results for lipids.

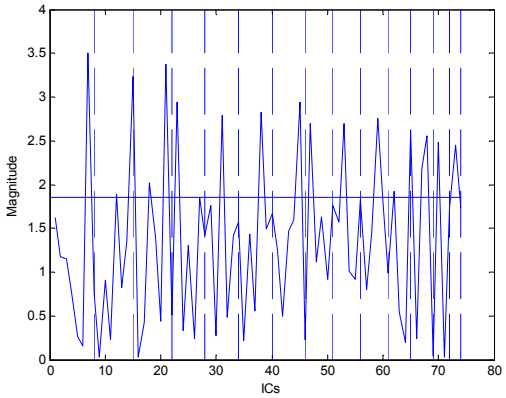

Fig. 7. Expanding window results for biomass.

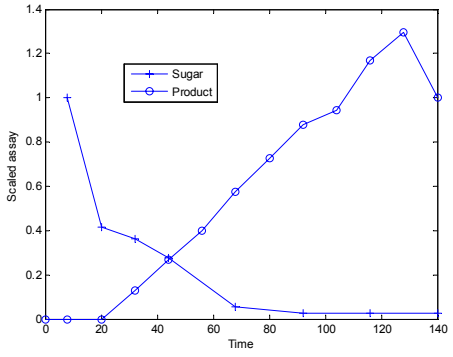

Fig. 8. Off-line assays of sugar and product

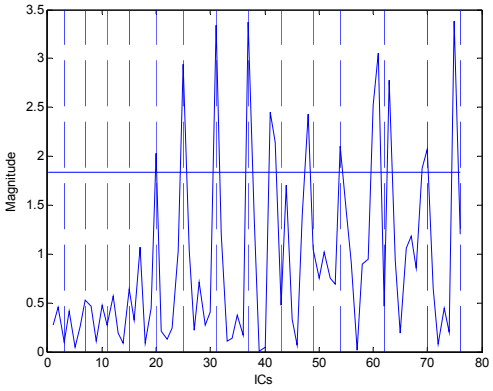

Fig. 9. Absorbance at 1224nm for all ICs over the course of batch 1

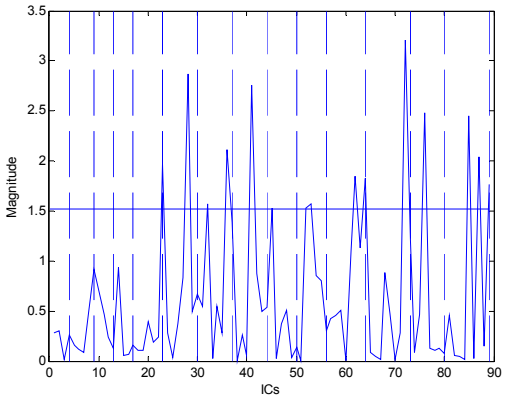

Fig.10. Absorbance at 1224nm for all ICs over the course of batch 2

To investigate whether ICA can be used to detect the appearance or disappearance of a compound, product and sugar are investigated. Product can be detected at 1224 nm. Currently, product assays are undertaken from around thirty hours as prior to this high substrate levels inhibit product formation. Fig. 8 shows the product and sugar profiles that occur over a typical batch as measured by off-line laboratory assay. It would thus be expected that the product spectra would not appear until thirty hours into the batch.

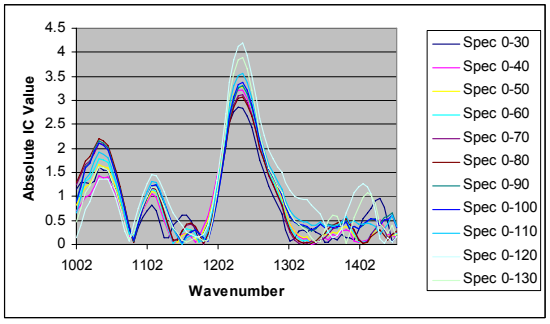

Fig. 11. Product spectra obtained via ICA from 0-30 hours onwards

Figs. 9 and 10 show the magnitude of absorbance at 1224 nm for all ICs for two batches. Both figures clearly indicate that a signal is not present until independent component 24 which lies in the 0-30 hour time window and subsequent windows show its presence. No spectra are obtained in the 0-10 and 0-20 hour windows. The spectra obtained by ICA from the time window 0-30 hours and subsequent windows are shown in Fig. 11. Again some variation in the peak height is observed due to changes in broth characteristics.

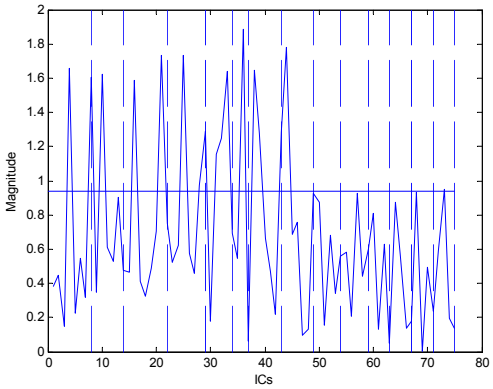

Fig. 12. Reverse extending window showing disappearance of sugar

The expanding window concept can be applied to a component that disappears towards the end of the batch if the window is expanded from the end of the batch to the start. For a batch that ends at 140 hours the windows would be 130-140 hours, 120-130 hours, etc. This concept is applied to detect the disappearance of sugar. Fig. 12 demonstrates that the sugar spectrum disappears at around 80 hours which is consistent with the off-line assay profile shown in Fig. 8.

3.3 Application of PCA to Independent Components from Multiple Batches

The preceding discussion focussed on the analysis of single batches to understand the onset and limitation of critical components. Building on this concept, the next step was to consider the application of ICA to multiple batches and investigate if the methodology could be utilised as a finger printing technique. One of the issues that needs to be addressed with respect to the comparison of the independent components generated from different batches is that they are not uniquely ordered. Thus to undertake a comparison, there is a need to address the lack of ordering with regard to the ICs.

One possible methodology to undertake this is that of principal component analysis (PCA). PCA is a multivariate statistical projection technique that identifies, in the data being analysed, the direction of greatest variability by defining a new set of latent variables that are a linear combination of the original variables, in this case IC's. Thus the ordering of the ICs will cease to be of any consequence and hence it

will be possible to extract information regarding the similarity of the batches.

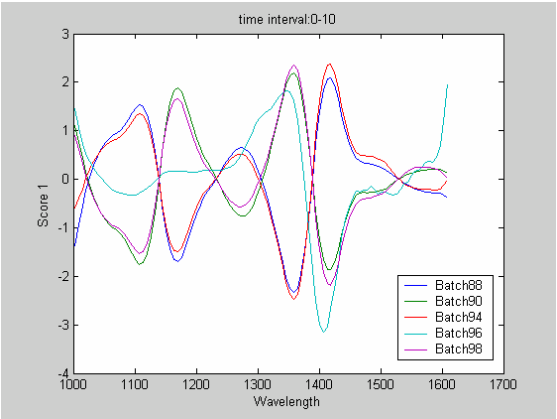

Fig. 13. Comparison of the IC's for all batches for principal component one

Fig. 13 illustrates the results from the application of PCA to the retained independent components, calculated for the five batches for the time period 0-10 hours. From these results it is evident that batch 96 differs to the other four batches. The interesting aspect of this is that through adopting this finger printing strategy, there is clear evidence in the first 10 hours of the batch that it exhibits different behaviour to the others and hence a more detailed study can be undertaken to isolate the cause of the difference and either corrective action can be undertaken or the batch terminated.

4. CONCLUDING COMMENTS

This paper has set out to demonstrate the capabilities of ICA for the interpretation of on-line NIR spectral data gathered from a fermentation process. The objective was not to construct spectral calibration models but to identify the presence or not of key compounds in the batch broth. This is important from a general operational perspective as compound limitations or excesses as a batch progresses can potentially cause product losses. In this study known compounds have been considered and it was found that it was possible to detect the appearance of a new IC relating to product concentration whereas other compounds were correctly identified as being present throughout the batch. Likewise, the limitation of a particular product, sugar was also identifiable from the analysis of the IC's. In the longer term, the benefits from applying ICA to NIR spectral signatures will be found in the detection of abnormal independent component profiles and since the IC is related to pure components, the peaks resulting will provide insight into the causes of deviation.

It was also observed that the IC derived pure component spectra show changes in peak magnitude as the batch progresses. If the ICs are utilised for compound detection then this is not a severe problem but if they are used as part of a calibration modelling procedure then such deviations will impact on prediction accuracy. In that case, the causes of variation such as scattering would need to be accounted for in the modelling algorithm.

5. ACKNOWLEDGEMENTS

The authors would like to acknowledge financial assistance provided by Newcastle University and the support of the CPACT consortium, in particular Spectroprobe and Clairet Scientific for the loan of the spectroscopic instrumentation. The authors are also indebted to Alison Dann from GSK (Worthing) and Alison Norden from CPACT Strathclyde for the experimental results.

6. REFERENCES

Back, A.D. and A.S. Weigend, (1997) "A first application of independent component analysis to extracting structure from stock returns", *Int. Journal of Neural Systems*, 8(4), 473-484.

Comon,P. (1994) Independent component analysis, a new concept? *Signal Process.*, 36, 287–314.

Gory, PA. (1990). General least-squares smoothing and differentiation by the convolution (Savitzky-Golay) method. Anal. Chem. 62:570-573.

Hyvärinen A. (1999). Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks* 10(3), 626-634, 1999.

Hyvärinen, A., Karhunen, J. and Oja, E. (2001) *Independent Component Analysis*, Toronto, Canada, John Wiley & Sons

Ladroue C., Howe F.A., Griffiths J.R., Tate A.R. (2002). Independent component analysis for automated decomposition of in vivo magnetic resonance spectra, *Magnetic Resonance in Medicine*, 50(4**)** , 69-703.

Lee S. and Batzoglou S. (2003) Application of independent component analysis to microarrays Genome Biology 2003, 4:R76

Scholz M., Gatzek S., Sterling A., Fiehn O. and Selbig J. (2004). Metabolite fingerprinting: detecting biological features by independent component analysis, *Bioinformatics*, 20(15), 2447–2454.

Vaidyanathan S, Harvey LH, McNeil B. (2001). Deconvolution of near-infrared spectral information for monitoring mycelial biomass and other key analytes in a submerged fungal bioprocess. *Analytica Chimica Acta*. 428, 41-59.

Vigario, R., Jousmaki, V., Hamalainen, M., Hari, R. and Oja, E. (1998). Independent component analysis for identification of artefacts in agnetoencephalographic recordings. *Advances in Neural Information Processing Systems* 10 229-235. MIT Press.