Recovering Independent Components from Shifted Data Using Fast Independent Component Analysis and Swarm Intelligence

CALEB RASCON,* BARRY LENNOX, and OGNJEN MARJANOVIC

Control Systems Centre, School of Electrical and Electronic Engineering, University of Manchester, P.O. Box 88, Sackvile Street Building, Manchester, M60 1QD UK

Frequency displacement, or spectral shift, is commonly observed in industrial spectral measurements. It can be caused by many factors such as sensor de-calibration or by external influences, which include changes in temperature. The presence of frequency displacement in spectral measurements can cause difficulties when statistical techniques, such as independent component analysis (ICA), are used to analyze it. Using simulated spectral measurements, this paper initially highlights the effect that frequency displacement has on ICA. A post-processing technique, employing particle swarm optimization (PSO), is then proposed that enables ICA to become robust to frequency displacement in spectral measurements. The capabilities of the proposed approach are illustrated using several simulated examples and using tablet data from a pharmaceutical application.

Index Headings: **Components; Shift; Swarm; Particle swarm optimization; PSO; Independent component analysis; ICA.**

INTRODUCTION

With the increased focus on process analytical techniques (PAT) in the pharmaceutical industry, there is an ever expanding use of spectroscopic instruments to provide important insight into the fundamental mechanisms driving particular processes.[1,2] A common application for spectral measurements in the pharmaceutical industry is identifying the concentrations of individual compounds that comprise a mixture. By knowing the concentration of individual compounds in a reactor, it becomes possible to track the progress of a reaction or the end-point of a process. A good example illustrating the benefits available through spectral analysis was documented by Szostak et al.,[3] who demonstrated how Raman spectral measurements could be used to identify the independent compounds in a pharmaceutical tablet. Further extension of this approach has enabled Raman measurements to be used to identify counterfeit medicines.[4]

Having identified the independent compounds in a mixture, the next stage of any analysis is typically to determine the concentrations of each of the independent compounds. This can be particularly important in the pharmaceutical industry; for example, Dyrby et al.[5] was able to identify the concentration of the active ingredient in a tablet, which could then be used to ensure that adequate mixing had occurred prior to tablet production.

Unfortunately, in many processing applications it is necessary to have *a priori* understanding of the spectral signature of the source components, which may not be available. To identify the independent components, or sources, in spectral or other types of data, an array of techniques have been proposed. Many of these techniques fall under the umbrella of blind source separation (BSS) methods.[6] One commonly applied BSS method is principal component analysis (PCA). This method is able to reduce the dimensionality of a problem by identifying directions of greatest variation in the data, with the imposed constraint that each of the directions are orthogonal. Each of these directions can be considered a "component" of the data, and the small-variance directions that are identified are typically considered to be "noise".[7] However, because only limited constraints are applied with this algorithm, the components that it identifies may have little physical meaning.[8]

A related method that imposes further constraints in its analysis is non-negative matrix factorization (NMF). As its name implies, all components that are identified are assumed to have no negative parts, which can make their physical representations easier to render.[8] However, it has been found that in certain circumstances it is difficult for the algorithm to converge to a global optimum, or even to converge at all.[9] One implementation of NMF is based on alternating least squares, which is considered to be part of a group of methods called self-modeling curve resolution methods (SMCR).[10] These methods have a similar objective to those of BSS, but their application is bound to the domain of the spectrum, as they were originally developed to be used in the field of chromatography. However, they are now beginning to gain popularity in the field of spectroscopy.[10] These methods frequently employ singular value decomposition or they consider the spectral intensity at specific locations that are common in all the sampled spectra. A disadvantage with these methods is that they require the number of components that are to be retrieved to be known *a priori*. Although this can be attempted in a variety of ways,[10] the presence of noise and other factors can make this procedure non-trivial.[10]

Independent component analysis (ICA) is an alternative approach to extracting meaningful components from a dataset. This technique has been used extensively in many different areas of science and engineering and its popularity is increasing.[11,12] Applications involving ICA have shown that it is a powerful and versatile method able to extract the independent components, or source signals, from spectra obtained from a variety of measuring devices such as near-infrared (NIR) and Raman instruments. The principle assumption that ICA makes is that all sources are independent of one another,[13] which is a reasonable assumption in many applications. It also provides an applicable objective function that, when optimized, identifies components that have real physical meaning.[14]

Much of the research conducted into using ICA to analyze spectral measurements has concentrated on its use in an off-line capacity. Moreover, the ability to analyze spectral data in real-time offers important benefits to industrial automation. In particular, the ability to analyze spectral measurements in real

Received 7 April 2009; accepted 19 June 2009.

* Author to whom correspondence should be sent. E-mail: Caleb.Rascon@postgrad.manchester.ac.uk.

0003-7028/09/6310-1142$2.00/0

time provides the possibility of using such measurements in a feedback control system. However, analyzing spectra in real time introduces a number of complex challenges. In this paper, the challenge of using misaligned or shifted spectra is addressed.

In an off-line setting, it may be reasonable to assume that the spectral data will not suffer from misalignments; however, this assumption is often not valid in real-time analysis. Inconsistencies related to spectral measurements are frequent as a result of poor sensor calibration and/or external influences.[15] In many cases these inconsistencies result in frequency displacement, or shift, in the measurements, and this has a significant effect on the ability of ICA to extract components with physical meaning from the spectra. Frequency displacement manifests itself by shifting the frequency location of important parts of a spectrum. The shifted parts of the spectrum can be the peaks that ICA uses to identify the different components that make up the spectrum. However, these shifts are not uniform, as each component inside the spectrum may shift independently from the rest.

Various approaches for tackling frequency shifts in spectral measurements have been presented. The most common approach is to recalibrate the sensors as required. Unfortunately, this approach is not ideal, as firstly it is not always obvious when a sensor requires recalibration and secondly, recalibration procedures can involve the use of expensive reference materials, as well as the loss of revenue because of the need to stop the plant to calibrate the sensor.[15] An alternative method for tackling this problem is to understand the external influences, such as temperature changes, affecting the measurements and to compensate for them. The influence of temperature on spectral measurements has been the subject of several research projects.[16–19] In particular, methods to identify the relationship between temperature and spectral measurements have been explored.[19] However, to build such a model, the temperature at which each measured spectrum was sampled needs to be known, together with the concentrations of each of the components in the measured spectrum. Whilst the former may be routinely measured, the latter is typically unknown.

Another approach is to align the data before applying ICA. Such a task is not trivial, as each component shifts independently. Artificially shifting a spectrum to align it with another, using the spectral features of one component as reference, will result in the misalignment of other components. Alternatively, more sophisticated aligning procedures may be used, such as dynamic time warping (DTW)[20] or alignment by fast Fourier transform (RAFFT or PAFFT).[21] These methods artificially distort the spectra in the alignment process, which introduces further problems in any further analysis, such as ICA, where it is assumed that the shape of a component remains consistent throughout the set of spectral samples. Applying these methods for pre-aligning will result in ICA identifying several components where it should have identified only one.

Given that it can be difficult to ensure that spectral measurements collected from a process do not suffer from shift, it is important that real-time and off-line analysis tools are able to cope with this effect. Unfortunately, the vast majority of research in this field has focused on ensuring that the measuring devices themselves do not produce shifted measurements and that if they do, techniques be developed for reducing or ideally eliminating this shift. Neither of these approaches ensures that the spectral measurements are free from shift before complex data analysis tools, such as ICA, are applied to them.

In this paper, a post-processing technique is developed that allows ICA to accurately identify the independent sources contained in spectral measurements despite the sources being shifted inside the data. When applying ICA to shifted spectral data, it is found that several more significant components are identified than would be otherwise. By combining these components together in a specific way, the proposed method is able to accurately identify the source components in the spectrum. Finding the most appropriate combination of the identified components can be formulated as a nonlinear optimization problem. There are many optimization algorithms that can be applied to solve this problem, one of the simplest being gradient descent. This method finds a close-to-optimum point in the solution space by following the "descent" flow around a starting point. Unfortunately, it is very easy for this algorithm to reach a local maximum, or minimum, and consider it to be the global optimum.[22] Genetic algorithms (GAs) are a popular optimization algorithm that have been designed to specifically search for a global maximum/minimum and avoid any local solutions.[23] A similar technique, and the one that is utilized in this paper, is particle swarm optimization (PSO). PSO applies social interactivity between the different areas of the solution space to arrive at the solution with significantly fewer iterations than a GA.[24] Reduced computation is a particularly important issue in this work, as the ultimate goal is to provide real-time analysis of spectral measurements.

In the following section of the paper, a brief overview of ICA and PSO is provided. Then, a simulation study is defined and the effect that frequency shifts have when applying ICA to this simulated data is illustrated. Afterwards, the proposed method is applied to a real pharmaceutical dataset.

THEORY

Independent Component Analysis. Independent component analysis (ICA) was conceived by Jutten and Herault in 1986[13] and formally defined by Comon in 1994.[25] ICA can be used to separate a multivariate measurement into a number of sub-components, or *sources*, and is therefore ideally suited to the analysis of spectral measurements of mixtures composed of independent compounds.

The independence between sources is used as the objective function within the ICA formulation. This is then optimized to identify the signals that are most independent of each other. The different ways that independence can be estimated and how it is maximized have resulted in several implementations of the ICA concept.[26–28] One popular implementation of ICA is FastICA,[6,29,30] introduced by Hyvärinen.[31] FastICA uses the amount of mutual information shared among the sources as a measure of independence, estimated using differential entropy or *negentropy*.[6]

The approach can be described by considering a series of mixtures $\mathbf{M}$, all of which are composed from a group of sources $\mathbf{C}$, which were mixed by a matrix $\mathbf{A}$, i.e., $\mathbf{M} = \mathbf{AC}$. An estimate of $\mathbf{C}$ (known as $\bar{\mathbf{C}}$) can be extracted from $\mathbf{M}$ by applying $\bar{\mathbf{C}} = \mathbf{BM}$, where $\mathbf{B}$ is a de-mixing matrix that ICA seeks to obtain.

To acquire $\mathbf{B}$, $\mathbf{M}$ is first mean-centered and whitened. The singular value decomposition of the covariance matrix $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ is then calculated, resulting in $\tilde{\mathbf{M}}$. Because $\tilde{\mathbf{M}}$ is orthogonal, the number of parameters that must be identified is reduced. The

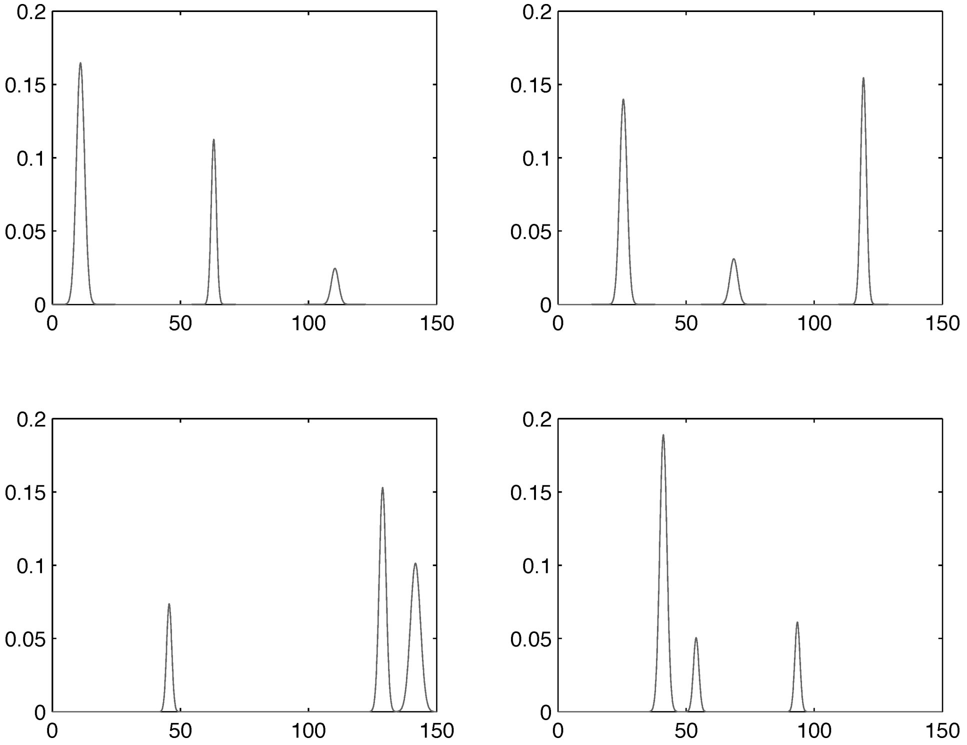

FIG. 1. Reference spectra randomly generated for use in experiments.

dimensionality of the problem can also be reduced in this stage by discarding small eigenvalues of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$.

After computing $\tilde{\mathbf{M}}$, $\mathbf{B}$ is found by using a fixed-point iteration scheme as defined below:

$$\mathbf{b}_i = \mathbf{E}\{\tilde{\mathbf{M}}\mathbf{G}(\mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{M}})\} - \mathbf{E}\left\{\mathbf{G}(\mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{M}})^{\mathrm{T}}\right\}\mathbf{b}_i \qquad (1)$$

$$\mathbf{b}_i = \mathbf{b}_i / ||\mathbf{b}_i|| \qquad (2)$$

repeat Eqs. 1 and 2 until convergence criteria is satisfied where $\mathbf{b}_i$ is the ith row of $\mathbf{B}$; $\mathbf{G}$ is the first-order derivative of a nonlinear function g that "does not grow too fast"[6] so it can converge at a minimal level of entropy. Equation 1 was derived by applying a constraint on the expected value of $\mathbf{b}_i^{\mathrm{T}}\tilde{\mathbf{M}}$ that satisfies Kuhn–Tucker conditions, making it possible to find its maximum value by a Newtonian method. The constraint applied is $||\mathbf{b}_i|| = 1$, which is met by applying Eq. 2. Both of these equations are applied until $\mathbf{b}_i$ converges.

The process repeats as many times as there are rows in $\tilde{\mathbf{M}}$. All the created $\mathbf{b}_i$s are then concatenated to form $\mathbf{B}$. However, more than one $\mathbf{b}_i$ may reach the same maximum, resulting in several estimates representing the same source. To avoid this, Eq. 3 is applied after each iteration of the fixed-point schemes to ensure that the rows in $\mathbf{B}$ are *non-correlated*. This method of de-correlation is referred to as *symmetrical* and is preferred for its equal weighting of all the $\mathbf{b}_i$s.

$$\mathbf{B} \leftarrow \mathbf{B} \cdot \left[(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{1/2}\right]^{-1} \qquad (3)$$

When $\mathbf{B}$ is calculated, the resulting $\bar{\mathbf{C}}$ will hold the estimated sources that, because of the small amount of mutual information between them, can be considered independent, hence the name independent components (ICs).

It is important to mention that throughout this article when using the term ICA, this refers to the FastICA implementation of this method.

Particle Swarm Optimization. Particle swarm optimization is a search algorithm that was introduced by James Kennedy and Russell Eberhart in 1995.[24] It is based on the inner social behavior of a flock or a school to find food.

A group of particles (or *swarm*) is randomly placed inside the solution space defined by an objective function. Each particle can "move" towards different locations in the solution space, and each location is graded by the objective function. Every particle is able to remember the best-graded location it has found, and makes it known to a pre-defined number of neighbors. During each iteration, the velocity of each particle is modified by considering the best-graded location found by the particle and the best one found by its neighbors, i.e.,

$$\begin{aligned} V_{x_i}(k+1) = V_{x_i}(k) &+ 2{*}r{*}(pbest_{x_i} - present_{x_i}) \\ &+ 2{*}r{*}(gbest_{x_i} - present_{x_i}) \end{aligned} \qquad (4)$$

where k is the iteration index, V_{x_i} is the velocity of the particle in the direction x_i, $pbest_{x_i}$ is the best-graded location in direction x_i found by the particle, $gbest_{x_i}$ is the best-graded location in direction x_i found by the neighbors of the particle, and $present_{x_i}$ is the current location of the particle in direction x_i. r is a stochastic factor that prevents several particles from being at the same location. It makes the particles "spread out" in an *area*, rather than focus on a single point, which improves significantly the chances of finding the true global optimum. All the V_{x_i}s of all the particles are modified according to Eq. 4 until the best-graded location found by the whole swarm converges or the maximum number of iterations is exceeded.

The PSO algorithm can incorporate the concept of a *time-decreasing inertia*,[32] which forcefully decreases velocities later

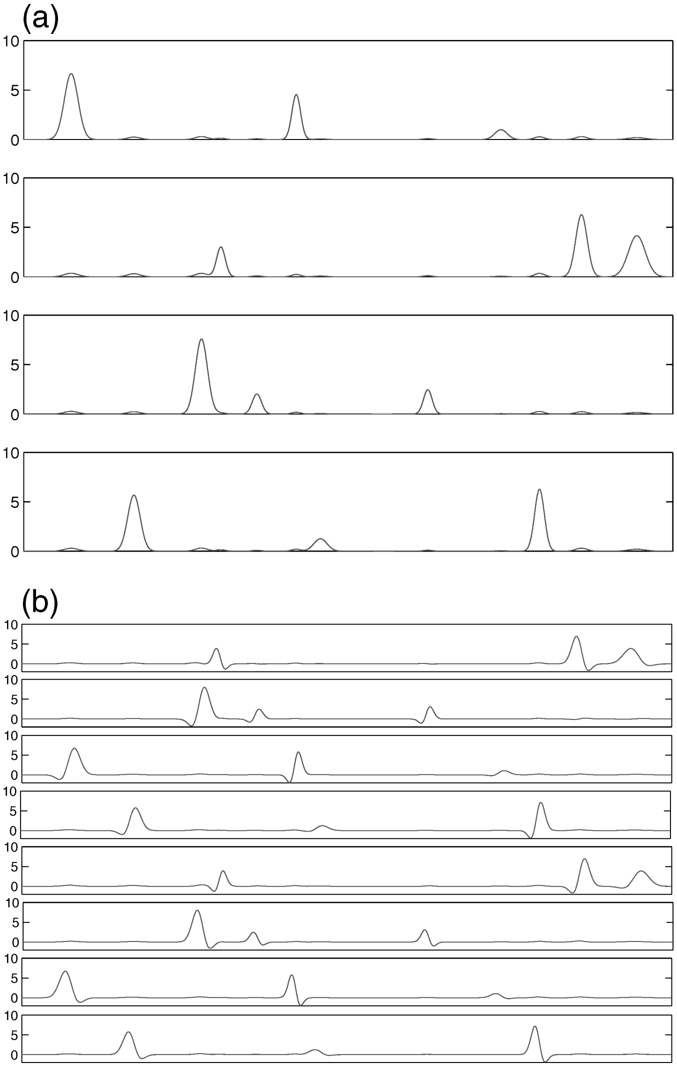

FIG. 2. ICs identified by ICA with and without applying shift. (**a**) ICs identified without shift. (**b**) ICs identified with a maximum shift of 1 *fp*.

in the search. This technique is an implementation of the temperature decrease in a simulated annealing search,[33] first introduced by Černý in 1985. Applying it to PSO results in an initial exploration of the whole solution space, pinpointing the area where the global optimum is suspected to be located. It then evolves into an exploitation of the area for the remainder of the search. It has been shown that using time-decreasing inertia in PSO provides faster and more accurate results than without.[32]

Application Study. To demonstrate the capabilities of the proposed method, artificial datasets that simulate the mixtures of four components are used. The reference spectra for these four components were randomly generated and are shown in Fig. 1.

The domain of the spectra is in Hertz and their resolution is 0.1 Hz per frequency point (*fp*). The structure of these spectra was defined such that they were consistent with data observed in the pharmaceutical and biomedical industry,[34] as well as other fields.[35,36]

Each dataset consisted of 100 samples, each containing a spectrum of a simulated mixture of the four reference spectra. Before being “mixed”, each spectrum was scaled by a factor randomly chosen between 0.2 and 1, simulating its concentration, and was optionally shifted by a random value, in terms of *fp*. A maximum shift value was given to each dataset (*max_shift*), defining the range of [–*max_shift*, *max_shift*], from which all the shift values applied to its samples were randomly chosen.

RESULTS AND DISCUSSION

Application of Independent Component Analysis. Two datasets, as defined in the previous section, were created; one was specified with a maximum shift value of 1 *fp* (0.1 Hz) while the other was left un-shifted. ICA was applied to both datasets; the number of components identified in each case was determined by the whitening process. The ICs identified from the non-shifted dataset are shown in Fig. 2a, and the ICs identified from the shifted dataset are shown in Fig. 2b.

Figure 2 shows that ICA is able to identify the four components in the non-shifted dataset, but it identifies eight components when using the shifted dataset. The trends displayed in Fig. 2b show that there are pairs of “similar” components, but the peaks are not located at the same frequencies and there are downward trends in the leading edges of the peaks that are very dissimilar to the shape of the sources.

Hyvärinen et al., pg. 1, wrote that “Actually, and perhaps surprisingly, it turns out that [to solve the ICA problem] it is enough to assume that [the sources] [...], *at each time instant t*, are statistically independent.”[6] Meaning that each source needs to be aligned properly in every mixture for it to be considered the same component. If not, ICA identifies “partial components” such as the ones shown in Fig. 2b. In this paper, this feature is referred to as *component division*.

These partial components are different from the sources not only in peak locations, but in peak shapes as well, so they cannot be considered as source estimates by themselves. If the maximum shift is increased, then further partial components for each source are identified. This is important as in more realistic situations the size of this shift is likely to vary continuously, and hence more partial components would be identified.

Proposed Post-Processing Algorithm. Figure 3a shows two related components that were identified when ICA was applied to the shifted dataset. Figure 3b shows the spectrum that results from simply adding these two spectra together. Figure 3c shows the reference spectrum that is most similar to the two spectra identified using ICA. These figures clearly show that by adding the two partial components together, an accurate approximation of the reference spectra is obtained.

The combined component illustrated in Fig. 3b is referred to in this paper as the estimated independent component (EIC) and is the estimate of the source to which the combined ICs are related. This relatively simple technique provides a feasible solution to identifying the source components in a spectrum affected by frequency shift. However, exhaustive testing has shown that when more partial components are identified, the simple addition of related components does not produce accurate approximation of the source spectra.

The post-processing algorithm proposed in this paper operates in three stages:

(1) Partial components that are related are identified and grouped.

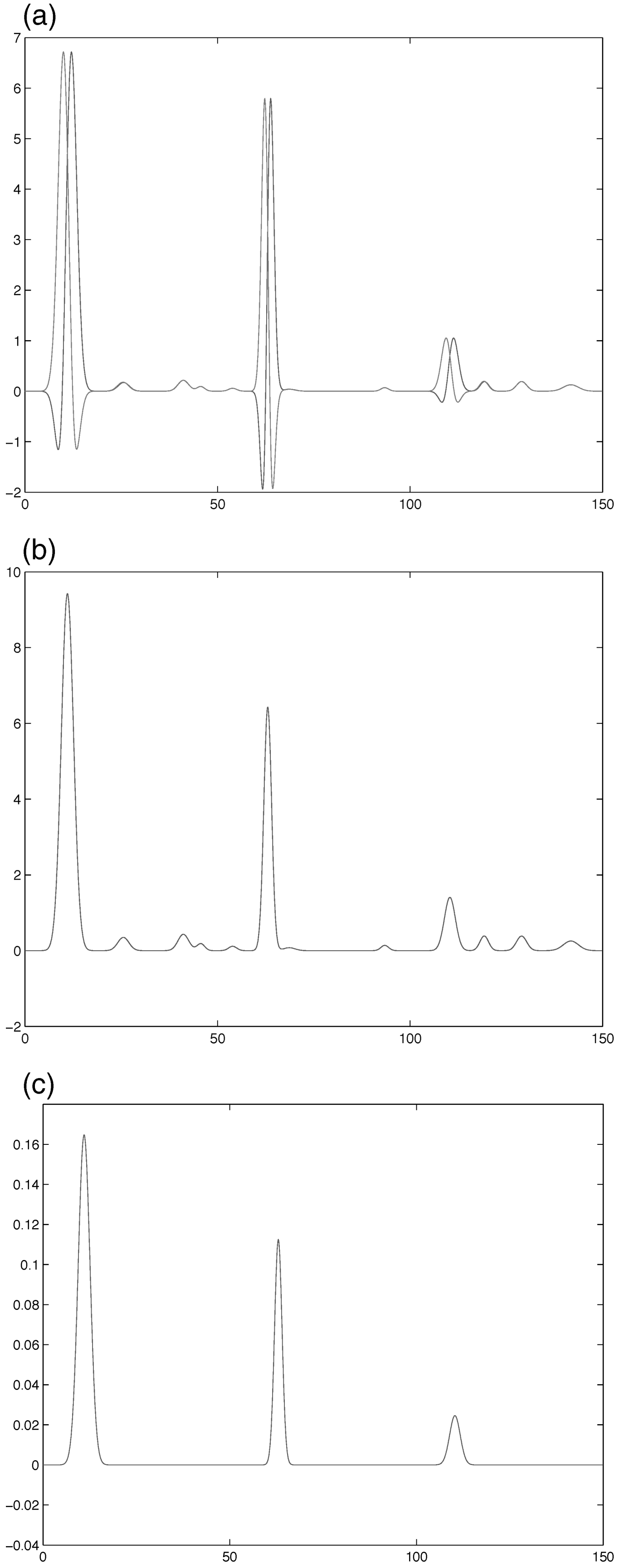

FIG. 3. Result of combining related ICs. (**a**) ICs found to be related. (**b**) Combination of the ICs. (**c**) Source to which the ICs are related.

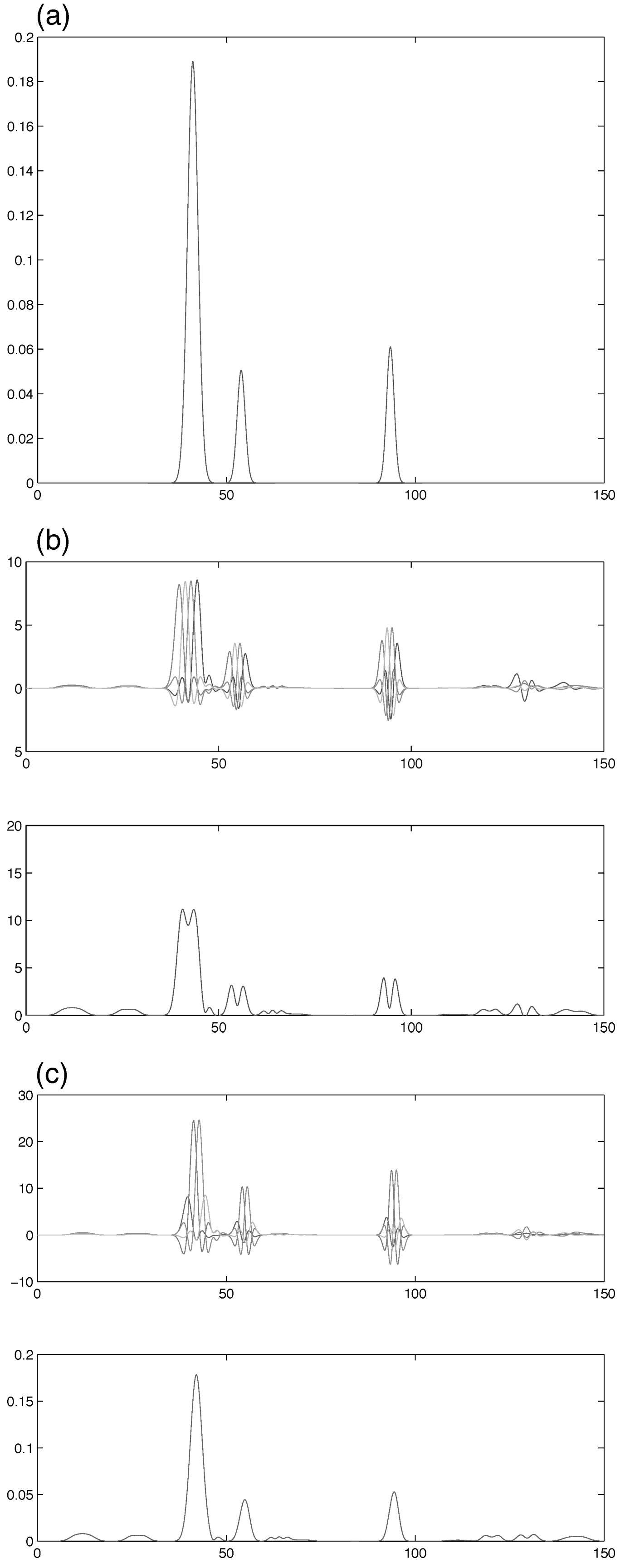

FIG. 4. Result of PSO search to correct the shape of the linear mix. (**a**) Source. (**b**) ICs from a 2 Hz maximum shift. (**c**) Shape-corrected EIC.

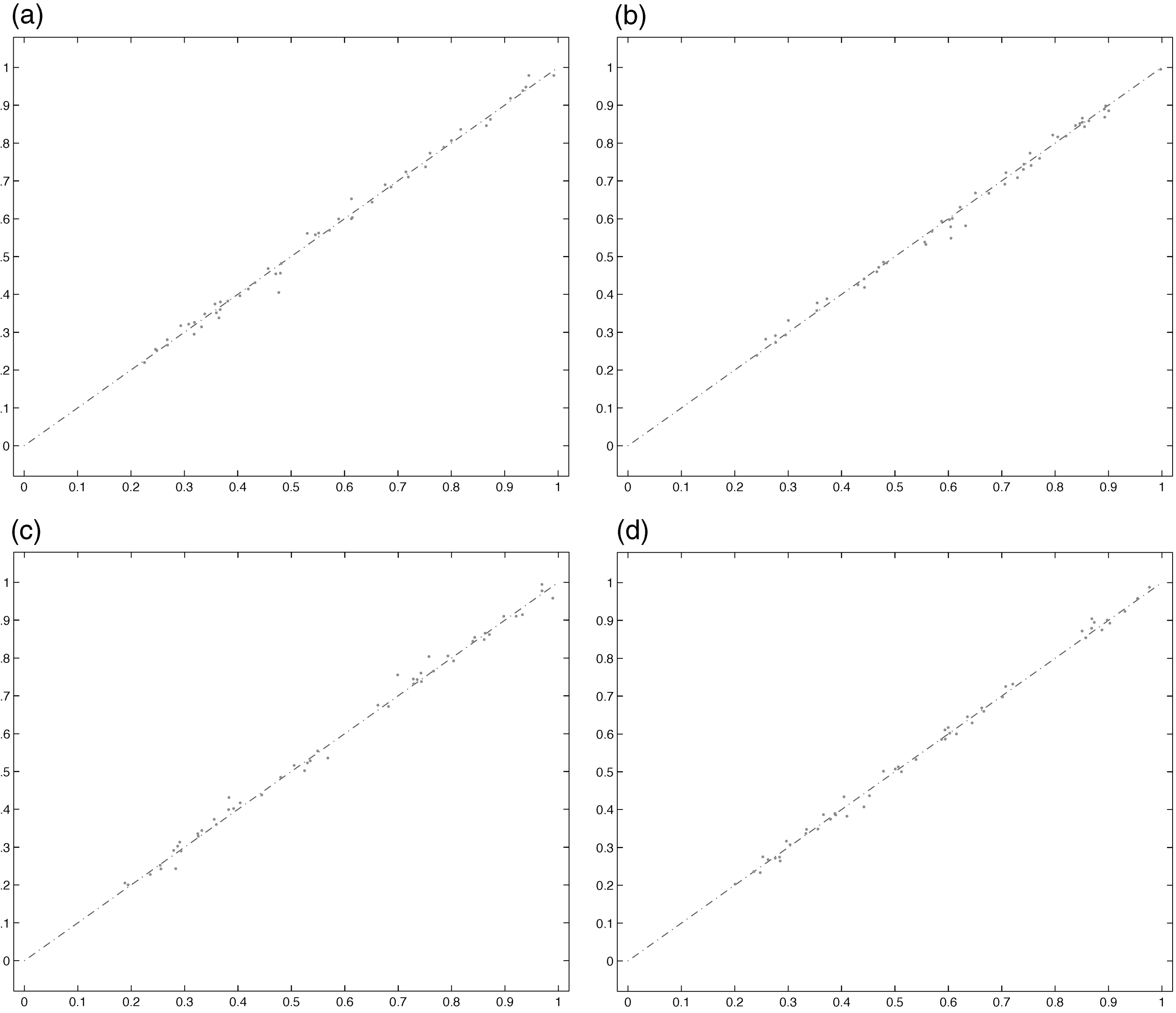

FIG. 5. PSO performance using the most similar EICs (0.2 Hz set). (**a**) Results for Comp. 1. (**b**) Results for Comp. 2. (**c**) Results for Comp. 3. (**d**) Results for Comp. 4.

(2) The grouped components are combined to produce an optimal EIC.
(3) Final processing of the EICs is undertaken to remove artifacts from other components.

Each of the these three stages is now described in detail.

Grouping Related Components. Two components that are "related" to each other will have a high correlation coefficient in an area near the origin in their resulting normalized cross-correlation vector (*NCCV*). In this example, an area of 40 *fp* (~8 Hz) with a cut-off value for the *NCCV* of 0.7 gave the best results in terms of finding which components were "related". However, studies suggested that the cut-off value of 0.7 was too strict using other datasets. A way to find a balanced cut-off value for a dataset is by applying all the values between 0.4 and 1 (with a step size of, say, 0.01) and recording the number of groups of ICs that were obtained for each value. Any values lower than 0.4 may give false positives of relation between ICs. The number of groups most frequently recorded was found to be a good estimate of the number of sources in the data, and any cut-off value that produced this number is appropriate. However, the cut-off value will not always be optimal, and some components may get left out. In such cases, user intervention may be necessary.

Combining Related Components. When only two components are found to be related, scaling them to be of the same height and adding them often provides a reasonable approximation to a source spectrum, as shown in Fig. 3. However, when more than two components are found to be related, it is not enough to equalize heights and combine them, as demonstrated in Fig. 4. In this example, four "grouped" components are identified and displayed in the upper graph of Fig. 4b. The resulting EIC displayed in the lower graph of Fig. 4b has a very different shape from the source shown in Fig. 4a.

To resolve this problem it is necessary to find the optimal combination of the scaling factors of the related components.

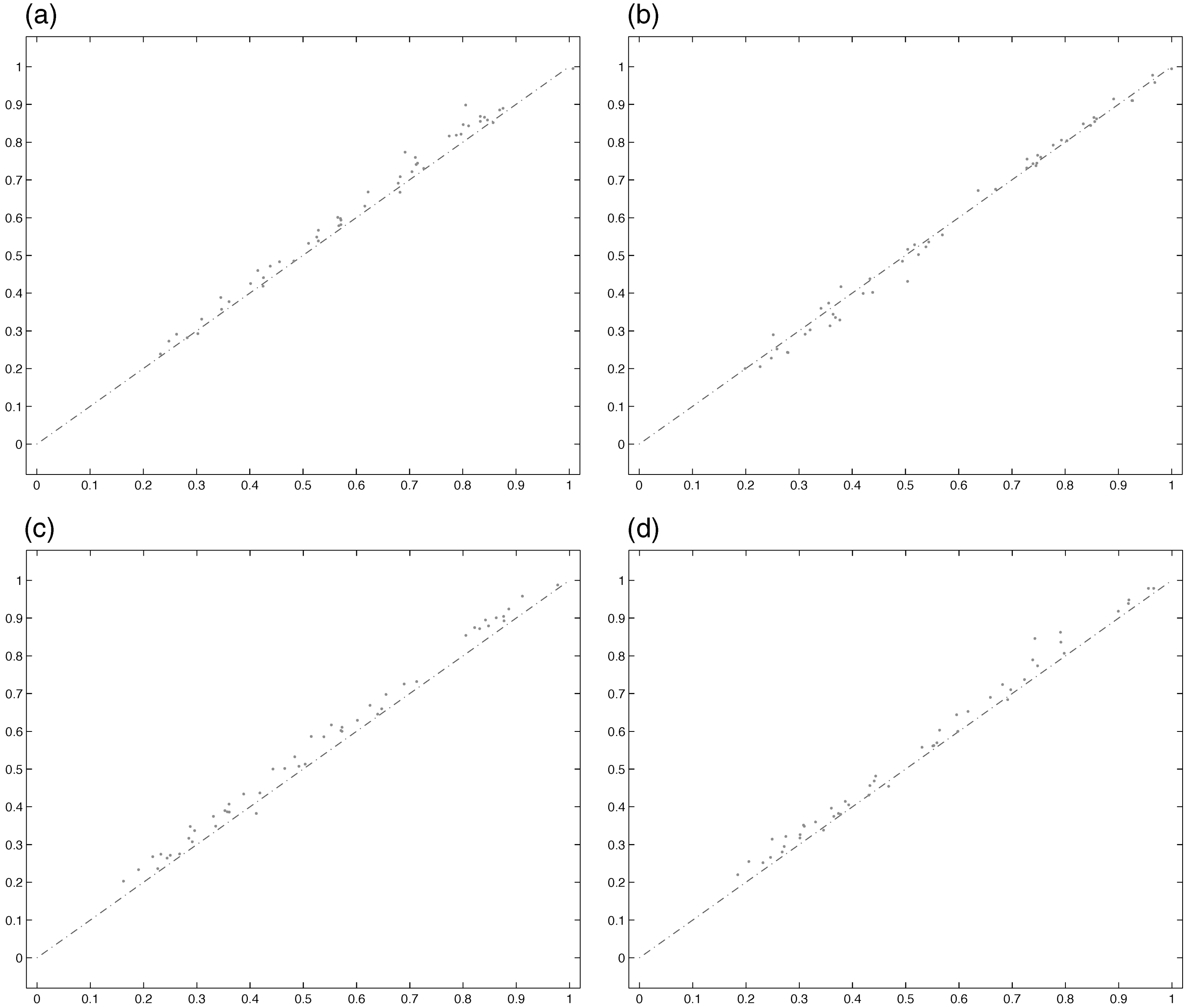

FIG. 6. PSO performance using the least similar EICs (1.1 Hz set). (**a**) Results for Comp. 1. (**b**) Results for Comp. 2. (**c**) Results for Comp. 3. (**d**) Results for Comp. 4.

For example, the upper graph of Fig. 4c shows the four rescaled related components that, when added together to provide the EIC in the lower graph of Fig. 4c, accurately describe the source in Fig. 4a. To find the optimal combination, PSO was applied. The optimal EIC was defined as:

$$EIC = \text{shift}\left(IC_1 + IC_n + \sum_{i=2}^{n-1} a_i IC_i,\ l\right) \tag{5}$$

and the following function was minimized:

$$P = \min[\text{Pearson}(EIC,\ data\_sample)] \tag{6}$$

where IC_i is the ith IC in the group, n is the number of ICs inside the group, the Pearson function is a measure of similarity based on the *Pearson product-moment coefficient*, and *data_sample* is a randomly chosen sample spectrum from the dataset. PSO aims to find the optimal combination of a_is (scaling factors applied to each IC_i). The range of the a_is was chosen empirically to be between 0.8 and 3, which provided good optimization speed without losing accuracy. The EIC is being compared to a data sample that may be shifted, so, to obtain an optimal fit, the EIC is artificially shifted an amount l that PSO also aims to find.

The Pearson coefficient was used as it only takes into account the similarity of the shapes of the spectra, regardless of their magnitudes. To force only one optimal combination to exist, the first and last ICs of the group (IC_1 and IC_n) remain unchanged throughout the search. If all the ICs vary, different combinations would exist that give the same measure of optimality.

Final De-Correlation. Figure 3b shows that when calculated, the EIC may contain artifacts from other components (e.g., the small peaks between 20–35 Hz), implying that the EICs are not completely de-correlated. To reduce this correlation, the following update to every EIC is applied until convergence is

TABLE I. The average similarity metric of each EIC set against the reference spectra.

Set	Corr.	Set	Corr.	Set	Corr.	Set	Corr.	Set	Corr.
0.1	0.9997	0.5	0.9945	0.9	0.9922	1.3	0.9924	1.7	0.9926
0.2	**0.9999**	0.6	0.9924	1.0	0.99924	1.4	0.9935	1.8	0.9917
0.3	0.9899	0.7	0.9930	1.1	**0.9871**	1.5	0.9937	1.9	0.9902
0.4	0.99942	0.8	0.9934	1.2	0.9923	1.6	0.9928	2.0	0.9882

reached:

$$EIC_i \leftarrow EIC_i - \left[EIC_j * \frac{EIC_i(f_{j_m})}{EIC_j(f_{j_m})}\right] \tag{7}$$

where EIC_i is the EIC to be updated; EIC_j is any other EIC; and f_{j_m} is the frequency location with the most energy in EIC_j.

Estimated Independent Components in Further Analysis. Once identified, the EICs can be used as reference spectra for further analysis, such as determining the concentrations of the various components in a measured spectrum. This can be achieved by finding a combination of scaling factors and shift values for each EIC, which, when combined, create a spectrum that is the closest in shape to the measured spectrum. To find such a combination, PSO can be applied to find an optimal:

$$C = \sum_{i=1}^{n} [c_i * \text{shift}(EIC_i, l_i)] \tag{8}$$

such that

$$P = \min(E) = \min\left\{\sum_{f}^{F}\left[C(f) - \frac{MS(f)}{||MS||}\right]^2\right\}^{1/2} \tag{9}$$

where *MS* is the measured spectrum; *C* is the spectrum created when applying the concentrations c_i and shift values l_i to their respective normalized EICs (EIC_i) and adding them; and *C*(*f*) and *MS*(*f*) are the energies located at frequency *f* in both spectra of size F. E is a measure of dissimilarity between *C* and the normalized measured spectra (*MS*/||MS||), based on the Euclidean distance between them, which has been found to give good results with this method. *P* is the minimum distance, which identifies the optimal combination of concentrations and shift values that best fit *MS*.

When the optimal combination is found, the estimated concentrations of the normalized version of the sources inside the measured spectra ($\hat{x}_i$) can be calculated by

$$\hat{x}_i = c_i * ||MS|| \tag{10}$$

Experiments and Results. For the following experiments, 20 different datasets were created, each having a different maximum shift value, which ranged between 1 *fp* (0.1 Hz) and 20 *fp* (2 Hz). The proposed post-processing technique was applied, and for every dataset, four EICs were identified, matching the number of reference spectra used.

The maximum correlation coefficient between the reference sources and their corresponding EICs shifted between −4 and +4 Hz, which was used as a measure of similarity. Table I provides the mean value of the similarity metric for each dataset, denoted by their maximum shift value. The bold numbers highlight the most and least similar sets of EICs. These values indicate that the proposed approach has successfully identified the four sources even in situations of severe shift.

To provide a measure of the accuracy with which PSO can estimate concentrations in measured spectra using these EICs, another dataset was generated with a maximum shift value of 1 Hz. In Figs. 5 and 6, the estimated and real concentrations are plotted, using the most and least similar EICs, respectively, and it can be seen that the concentrations are well estimated. The mean square error was 2.9707×10^{-04} and 9.9480×10^{-04} for the 0.2 Hz and 1.1 Hz EIC sets, respectively. These results

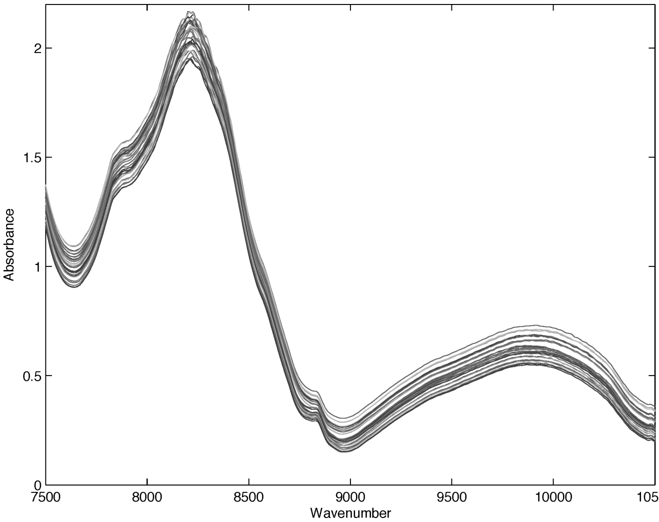

FIG. 7. Spectral data of pharmaceutical tablets.

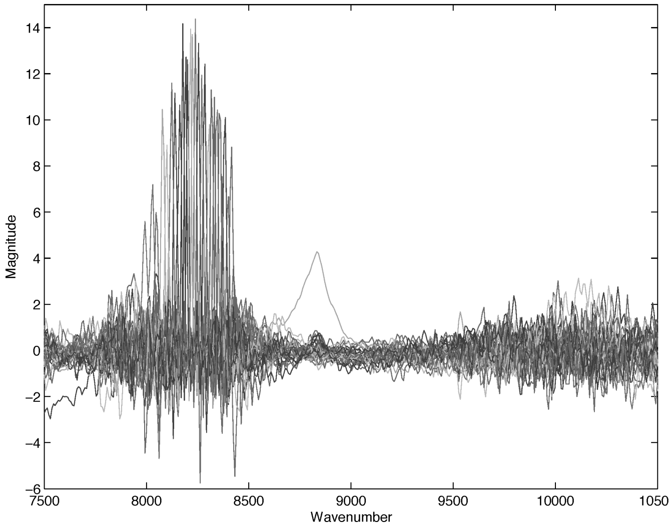

FIG. 8. ICs that were found by ICA in spectral data.

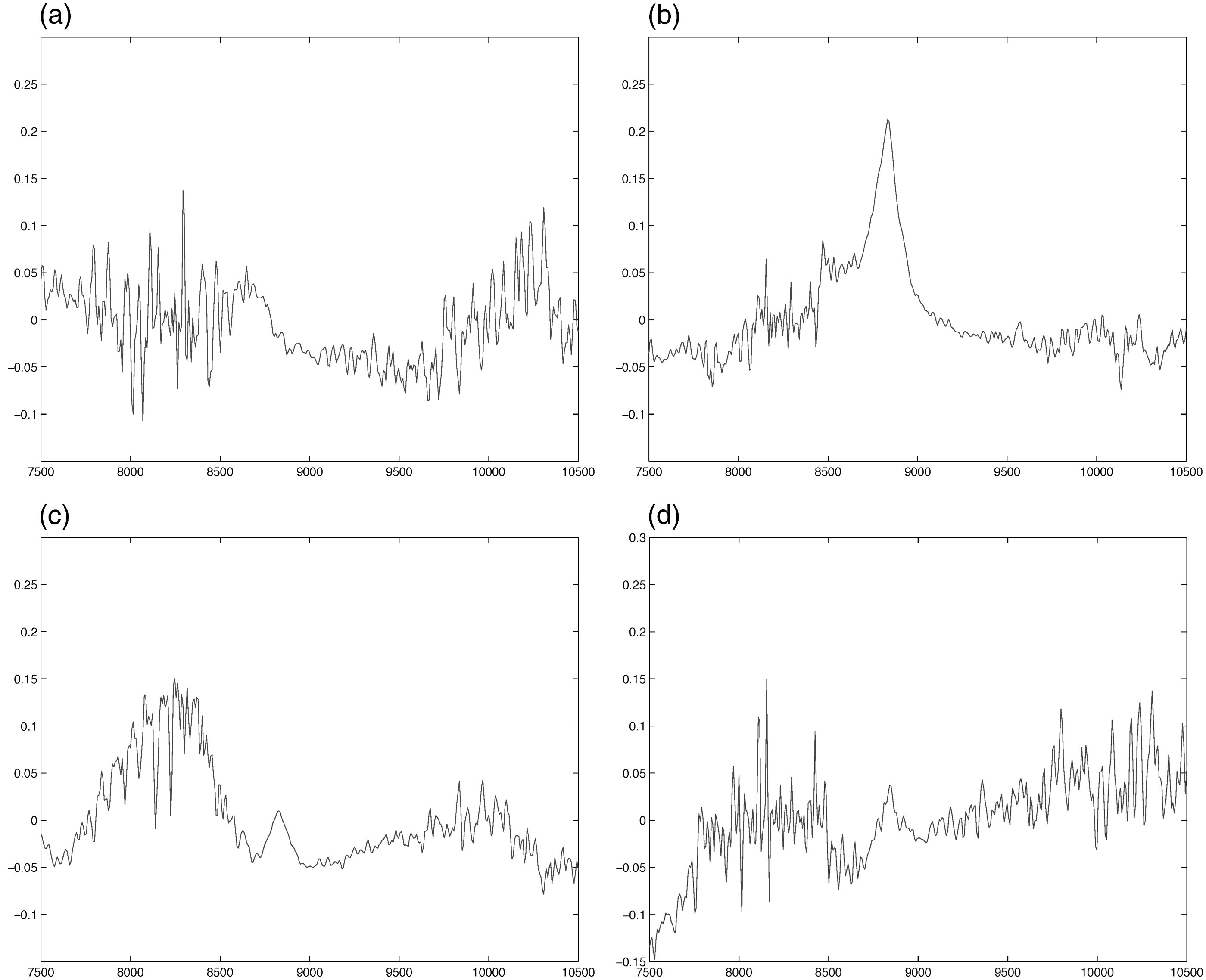

FIG. 9. EICs obtained from the results in Fig. 8. (**a**) EIC 1. (**b**) EIC 2. (**c**) EIC 3. (**d**) EIC 4.

highlight the accuracy with which PSO can estimate the concentration of each identified component in the mixtures.

CASE STUDY

In this section the ability of the proposed technique to compensate for shift in a real set of spectral data was investigated. The dataset that was used in this study is part of a public database consisting of 310 NIR spectra, made available by Dyrby et al.,[5] sampled from pharmaceutical tablets of Escitalopram®. It is composed of several batches, differing by the size of the tablet (5, 10, 15, and 20 mg), and each batch has three sub-batches that differ in production scale (full scale, pilot scale, and laboratory scale).

The mixture inside the tablets is composed of an active ingredient and several excipients, such as mycrocrystalline cellulose (dominant), magnesium stearate, and talc. In this section, the results when analyzing the batch of laboratory-scaled 5 mg tablets is presented. The results obtained using this dataset were comparable to those from the other datasets. Figure 7 shows the spectra from this batch of measurements.

Detailed examination of the spectra in Fig. 7 suggests that there is some frequency shift in the data, specifically in the area between 8000 and 8500 cm^{-1}, as well as the incidence of a small amount of noise. When ICA is applied to this data, 30 components were identified (shown in Fig. 8). Considering that there are 30 spectra in the dataset, it can be deduced that ICA is unable to reduce the observations to their unique components. This result illustrates the sensitivity of ICA to frequency shift.

The proposed shift compensation algorithm was applied to the dataset, and the resulting components are shown in Fig. 9. Details provided by Dyrby et al.[5] suggest that the active ingredient has an important peak near 8830 cm^{-1}, and visual comparison of Fig. 9b to the reference spectrum provided in the original work indicates that the component in Fig. 9b is similar to the spectral signature of the active ingredient. It is suggested[5] that the dominant excipient, mycrocrystalline cellulose, has a prominent peak near 8200 cm^{-1}, indicating that the component in Fig. 9c is a good candidate for this material. Information on the other materials in the tablet is not available.

It is important to note that the spectral signature of the active ingredient did not suffer from shift in the dataset. However, the severe shift in the other components made it difficult to correctly identify it using ICA alone. This implies that even from the same batch, seemingly irrelevant components (such as excipients in pharmaceutical tablets) may cause difficulties when identifying the active ingredients.

CONCLUSION

This paper has demonstrated that FastICA is not robust when applied to spectral data that suffers from frequency displacement. However, by using a post-processing algorithm it is possible to recover the shape of the spectra of the source components. The proposed approach was successfully applied to both simulated examples and to pharmaceutical NIR data.

In each study the proposed algorithm was able to recover a set of components that were very similar to the expected components. Further analysis showed that the accuracy with which the reference components could be identified from the data was such that the concentration of each of the components in the data could be identified with a very high level of accuracy.

1. W. W. Blaser, R. A. Bredeweg, R. S. Harner, M. A. LaPack, A. Leugers, D. P. Martin, R. J. Pell, J. Workman, and L. G. Wright, Anal. Chem. **67,** 47 (1995).
2. T. Togkalidou, M. Fujiwara, S. Patel, and R. D. Braatz, J. Cryst. Growth **231,** 534 (2001).
3. R. Szostak and S. Mazurek, Analyst **127,** 144 (2002).
4. M. de Veij, A. Deneckere, P. Vandenabeele, D. de Kaste, and L. Moens, J. Pharm. Biomed. Anal. **46,** 303 (2008).
5. M. Dyrby, S. Engelsen, and L. Nørgaard, Appl. Spectrosc. **56,** 579 (2002).
6. A. Hyvärinen and E. Oja, Neural Networks **13,** 411 (2000).
7. M. Ringnér, Nature Biotechnol. **26,** 303 (2008).
8. D. D. Lee and H. S. Seung, Nature (London) **401,** 788 (1999).
9. M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons, Comput. Stat. Data Anal. **52,** 155 (2007).
10. J.-H. Jiang, Y. Liang, and Y. Ozaki, Chemom. Intell. Lab. Syst. **71,** 1 (2004).
11. C. Ladroue, F. A. Howe, J. R. Griffiths, and A. R. Tate, Magn. Reson. Med. **50,** 697 (2003).
12. H. Malika, A. Khokhar, and R. Ansari, Proceedings of the 5th ACM Workshop on Digital Rights Management, 102 (2005).
13. C. Jutten and J. Herault, Signal Proc. **24,** 1 (1991).
14. M. Bressan, D. Guillamet, and J. Vitria, IEEE Conference on Computer Vision & Pattern Recognition **1,** 1063 (2001).
15. M. Zuppa, C. Distante, K. C. Persaud, and P. Siciliano, Sens. Actuators, B **120,** 411 (2007).
16. M. A. Czarnecki, Y. Ozaki, M. Suzuki, and M. Iwahashi, Appl. Spectrosc. **47,** 2157 (1993).
17. K. H. Hazen, M. A. Arnold, and G. W. Small, Appl. Spectrosc. **48,** 477 (1994).
18. T. Iwata, J. Koshoubu, C. Jin, and Y. Okubo, Appl. Spectrosc. **51,** 1269 (1997).
19. Z.-P. Chen, J. Morris, and E. Martin, Anal. Chem. **77,** 1376 (2005).
20. H. Sakoe and S. Chiba, IEEE Trans. Acoustics, Speech Signal Proc. **26,** 43 (1978).
21. J. W. H. Wong, C. Durante, and H. M. Cartwright, Anal. Chem. **77,** 5655 (2005).
22. S. J. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall, New York, 2002), 2nd ed., p. 115.
23. D. Barash, Ph.D. Thesis, University of California, California (1999).
24. J. Kennedy and R. Eberhart, Proc. IEEE Int. Conf. Neural Networks **IV,** 1942 (1995).
25. P. Comon, Signal Proc. **36,** 287 (1994).
26. A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (John Wiley and Sons, New York, Chichester, 2001).
27. S. Roberts and R. Everson, *Independent Component Analysis: Principles and Practice* (Cambridge University Press, Cambridge, UK, 2001).
28. J. V. Stone, *Independent Component Analysis: A Tutorial Introduction* (MIT Press, Cambridge, MA, London, 2004).
29. K. Ichige, M. Imai, and H. Arai, 14th Workshop on Statistical Signal Processing, 546 (2007).
30. Y. Tie and M. Sahin, Neural Eng. **2,** 90 (2005).
31. A. Hyvärinen, IEEE Trans. Neural Networks **10,** 626 (1999).
32. Y. Shi and R. C. Eberhart, Proceedings of the 7th International Conference on Evolutionary Programming VII, 591 (1998).
33. V. Černý, J. Optimization Theory Appl. **45,** 41 (1985).
34. W. P. Findlay and D. E. Bugay, J. Pharm. Biomed. Anal. **16,** 921 (1998).
35. J. C. Brown, J. Acoust. Soc. Am. **92,** 1394 (1992).
36. M. Castanys, M. J. Soneira, and R. Perez-Pueyo, Laser Chem., 11 (2006).