35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Fast and Robust Extraction of Reliable Protein Signal Profiles from Mass Spectrometry Data by Introducing the Concept of Single Channel ICA with Statistical Offset Correction

Mavuduru Neehar and Amit Acharyya

Abstract— **In this paper, we introduce the concept of Single Channel Independent Component Analysis (SCICA) for fast extraction of protein profiles from the mass spectra data. Subsequently we propose one offset-correction scheme on the basis of the statistical data analysis of the SCICA-based estimated protein profiles to ensure robustness of the proposed algorithm. The proposed method is also validated rigorously against the simulated data. Such concept, to the best of our knowledge is proposed for the first time in this context of protein profiling and we envisage that the proposed concept will find potential applications in bio-marker discovery especially for cancers.**

I. Introduction

In this paper, we propose to introduce, for the first time to the best of our knowledge, the concept of Single Channel Independent Component Analysis (SCICA) for fast extraction of protein profiles from the mass spectra data. Subsequently we propose one offset-correction scheme on the basis of the statistical data analysis of the SCICA-based estimated protein profiles to ensure robustness of the proposed algorithm.

ICA has recently been introduced in [1] to be used for the processing of proteomic signals particularly in protein profiling from the mixed mass spectra data. Because of the mixing of several protein peaks of different amplitudes corresponding to the abundance of various proteins, and the contaminations by several biological and physical artifacts [2], such mass-spectra data always present complex features. Due to the presence of these disturbances, very sensitive and accurate peak-detection methodologies, capable for robust separation of protein signals, are required. Several such methodologies have been reported in the literature for analyzing SELDI and MALDI-TOF data [1-6], however, the problem of robust protein peak detection has not been completely resolved. This problem seriously limits the development of reliable proteomics tools for biomarker discovery and early disease diagnosis [1]. Although application of ICA in this domain, as proposed in [1], has been proved to be much more beneficial than other existing methods like wavelet transform techniques [3], ICA has two main potential limitations in this context. Firstly, it requires the number of observations to be equal to the number of independent protein profiles. This means if there are n number of protein profiles in the mass spectra, n number of observations are necessary to separate these using ICA. Therefore n number of observation profiles has to be designed through n number of experiments by different calibrations which, in turn, consumes significant amount of time. Secondly, in this method of separation using ICA, the number of protein profiles is needed to be known beforehand. However, it is an unrealistic assumption leading to a sub-optimal solution and unnecessarily consumes significant amount of computational resources and time. These afore-mentioned problems motivate us to investigate the applicability of SCICA in this domain which will be discussed in Section-II followed by its validation in Section III. Subsequently we propose a statistical data analysis based offset-correction scheme to bring robustness in the protein profiling. To the best of our knowledge, this is the first of its kind work where SCICA is being explored to extract protein profiles from the mass spectra.

MN and AA are with the Department of Electrical Engineering, Indian Institute of Technology (IIT), Hyderabad, Andhra Pradesh, India – 502205. (e-mail: ee09b017@iith.ac.in, amit_acharyya@ iith.ac.in).

This work is supported by the DIT, India under the Cyber Physical Systems Innovation Hub under Grant number: 13(6)/2010-CC&BT , Dated 01.03.11.

II. Mehtods

A. Existing state-of-the art ICA based separation

ICA based signal separation method is the state-of-the art for protein profiling and was introduced in this domain by Mantiniet. al. in [1]. It assumes that the spectra (**X**) obtained from the MALDI-TOF devices as the linear combination of individual protein profiles (**S**). This matrix **X** is the input to the ICA algorithm. After the application of ICA, the outputs are two matrices: the first is the matrix **S** which contain the IC (Independent components) waveforms with the component IDs in rows and the intensities corresponding to the m/z values in columns; the second is the matrix **A** of the IC amplitudes with the component IDs in columns and the spectrum weights in rows.

B. Introducing SCICA in Protein Profiling

If the number of observations is less than the number of the independent sources mixed then this constitutes an under-determined blind source separation problem which can be solved using underdetermined ICA [7-9]. Single Channel Independent Component Analysis is a special case of an undetermined ICA [7-9]. It has only one observation which serves as the input for the SCICA algorithm and it estimates the contribution of each source present in the observation. The single channel ICA in essence is the instance of the multi-dimensional ICA (MICA) applied to vectors of delayed samples. This means that multiple ICA components

may be associated with a single independent source. A signal can be represented as the linear superposition: $\boldsymbol{x} = \sum_i s_i \boldsymbol{a_i} = \boldsymbol{As}$, where $\boldsymbol{s}$ is a column vector with independent sources $s_1, s_2, \ldots, s_n$ as its elements and $\boldsymbol{A}$ is a mixing matrix with its columns, $\boldsymbol{a_i}$ (also called weights) chosen such that they form a basis in the signal space [9]. This representation of a signal can be used to show the analogy with the problem of protein profiling from the mass spectra data. The vectors $\mathbf{s_i}$ can be assumed to be our sources (protein profiles) that are mixed, vectors $\mathbf{a_i}$ contain the information about how the sources are mixed in each observation and the matrix **X** represents our observation vectors which are the mass spectra (SELDI/MALDI-TOF) data.

For an ICA problem, the number rows of the **X** have to be greater than or equal to the number of rows of **S.** The ICA yields the unmixing matrix **W,** which can be used to extract the sources form the observations. So, the sources can be obtained back form the observations using the relation **S = W.X,** where $\mathbf{W = A^{-1}}$. In SCICA problem, the **x** is not a set of observations, but is only one vector. The matrix **X** has to be formed from this single observation vector. Then the normal ICA algorithm will be applied to find the mixing and unmixing matrices. To obtain the 'multi-channel representation of a single data channel, a series of delay vectors are to be generated to form a matrix of delays [9]. This matrix can also be called as embedding matrix. Assuming a single channel data with n elements: $\{x_i\}_{i=1,\ldots,n}$, then delayed vectors in the constructed matrix are given as $V_k = \{x_k, x_{k+1}, \ldots\ldots, x_{k+m-1}\}$. The delay matrix **X** is formed by obtaining V_k for successive values of *k*, and combining these to form:

$$X = \begin{bmatrix} x_t & x_{t+\tau} & \cdots & x_{t+m\tau} \\ x_{t+\tau} & x_{t+2\tau} & \cdots & x_{t+(m-1)\tau} \\ \vdots & \vdots & \ddots & \vdots \end{bmatrix} \qquad (1)$$

where τ is the lag term, and m is the number of lags (or the embedding dimensions. The number of lags m needs to be big enough to capture the information content necessary). Considering the value of τ to be equal to 1, we obtain a multi-channel representation of the single observation vector. Now, the algorithm proceeds in the following way:

1. Whiten the signal (with possible dim. reduction) using PCA, DFT or DCT.
2. Apply an ICA algorithm to find the mixing and unmixing matrices M and W.
3. Calculate the magnitude transfer functions of the basis vectors $a_i(t)$ and cluster into groups γ_p, p=1,…,*C* using k-means (or another clustering algorithm).
4. Calculate the shift-invariant separation and reconstruction filters, f_p, for each source using the equation:

$$f_p(t) = \frac{1}{N}\sum_{i \in \gamma_p} a_i(-t) * w_i(t) \qquad (2)$$

5. Now, pass the original time series through these filters to find the contribution of each source in the observation signal.

C. Separation using SCICA for Protein Profiling

Since the peaks present in the mass-spectra data have morphological resemblance to the frequency spectrum of any time domain sinusoids, without any loss of generality, we can assume that the observed. mass-spectra are in the frequency domain. This time domain signal is the input to the SCICA. The contributing signals are sinusoids with different frequencies (that correspond to the peaks in the spectrum) and it is also to be noted that these signals are spectrally disjoint and independent. So, these conditions realistically satisfy the algorithmic requirements of the SCICA algorithm. Hence the protein profiles can be separated from the original signal using SCICA algorithm. Assuming X_{syn} is the mass spectra and as discussed above, assuming it is in the frequency domain, inverse Fourier transform is applied on it to obtain Y as follows: $Y = f^{-1}\{X_{syn}\}$ which is then used as the input to the SCICA and is split into delay vectors those are then combined to form the delay or embedding matrix **X** as shown in (1). Whitening (with possible dimensionality reduction) is then performed on **X** before applying the ICA algorithm. Due to its higher convergence speed and superior accuracy with inbuilt whitening unit, FastICA is used to obtain the outputs of **S** with independent components along with **A** and **W**. Assuming the independent protein profiles are spectrally disjoint, *k-means* clustering algorithm is then performed on the basis vectors of **A** because multiple independent components can be associated with the same independent sources. Now each source can be calculated using (2). To obtain the contribution of each source in the observation domain the original time series data **Y** need to be convolved with the filters formed. This yields the independent sources in the time-domain as follows: $s_i = Y * f_{pi}$, where s_i represents the contribution of i^{th} source in the observation and f_{pi} represents the corresponding shift-invariant filter which is calculated after clustering the basis vectors. The procedure for extracting the protein profile using the SCICA can be summarized as follows:

1. Convert the frequency domain signal into time domain using inverse Fourier transform.
2. Split this time series data to form the delay vectors.
3. Construct the embedding matrix as described in the SCICA background section.
4. Apply the FastICA technique to learn the mixing and unmixing matrices.
5. Find the Transfer functions of the basis vectors $a_i(t)$, and cluster them into groups.
6. Calculate the shift-invariant separation and reconstruction filters, f_p, for each source using the equation (6).
7. Convolve the time series signal with these filters to find the contribution of the each source.
8. Apply the Fourier transform on these obtained independent source contributions to get the corresponding protein profiles.

III. Experiments, Results and Discussion

A. Experimental Set-up and Simulations

Adopting the method proposed in [1] and using the equation as given in [10] mass spectra peaks are generated:

$$x(z) = \frac{A}{\tau}\exp\left(\frac{\sigma^2}{2\tau^2} - \frac{z-z_p}{\tau}\right)\int_{-\infty}^{h}\frac{1}{2\Pi}\exp(\frac{-t^2}{2})$$

where z is the mass/charge (m/z) value, A is the area of the peak, $\boldsymbol{\tau}$ is the time constant of the exponential decay, σ controls the tailing of the peak, z_p determines the position of the peak on the m/z axis, the ratio $\boldsymbol{\tau}$ / σ is a measure of its asymmetry, and $h = \frac{z-z_p}{\sigma} - \frac{\sigma}{\tau}$.

A synthetic MALDI-TOF mass spectra dataset X_{syn} was prepared by means of a linear mixing of the 10 protein profiles which we will consider for our study. The amplitudes and positions of peaks vary from 600 to 1000 and 100 to 1000 respectively. The initial data set consists of 2100 samples. The delay dimension is set to 100 and embedding matrix of size 100X900 was constructed using the delay vectors. The MATLAB FastICA package downloaded from http://www.cis.hut.fi/projects/ica/fastica/ was used. FastICA contains the whitening step before the application of the ICA technique and the mean of the data will be added back to the ICs after their calculation. The simulation was ran on an Intel Core-i5 processor @ 2.40GHz with 4.00GB RAM. Inverse Fast Fourier transform *(ifft)* command present in the MATLAB is used to convert the original signal into the time-domain. After splitting this data into delay vectors to form a matrix, the FastICA is applied. After calculating the filters, the contribution of the sources in the observation is found by convolving the time-series signal with each of these filters separately. The MATLAB command *conv* is used for performing this operation. Now, the Fast Fourier transform (*fft*) which is another inbuilt command in the MATLAB is used to convert these independent sources back into the frequency domain (which is in the same domain as the original input).

As a part of our preliminary research and to prove our proposed concept, we assumed that the mass-spectra data (mixture) contain 10 individual protein profiles with proteins arbitrarily positioned at 127, 161, 243, 394, 470, 540, 643, 730, 890 and 964 as shown in Fig. 1 (a), (e), (i), (m). However, it is to be noted that we tried to depict the practicality of any protein profiling system where proteins can be up- or down-regulated depending upon various diseased conditions. Considering Fig. 1(a) represents the protein profiles depicting normal condition with four protein peaks are of high amplitudes (top row, now onwards refer to as strong proteins), two peaks are moderate amplitudes (middle row, now onwards moderate proteins) and four others have little amplitude (bottom row, now onwards weak proteins). Under diseased conditions, we simulated three more cases: (A) four strong proteins are down-regulated to moderate and weak proteins (Fig. 1(e)), (B) one moderate protein is up-regulated to strong protein and the other one is down-regulated to weak protein (Fig. 1(i)) and (C) four weak proteins are up-regulated to strong proteins and moderate proteins (Fig. 1(m)). All these afore-mentioned four cases including the normal condition, are mixed linearly with random vectors to simulate the mass-spectra data (mixed) as shown in Fig. 1(b),(f),(j),(n) on which SCICA is applied.

B. *Results and Discussion*

It is important to notice that the extracted protein peaks using SCICA are positioned at 126, 160, 242, 393, 469, 539, 642, 729, 889 and 963 as shown in Fig. 1(c), (g), (k), (o). It is to be noted that there is a consistent lag in the protein peak position in the extracted signal from the original protein profiles (e.g. 126 instead of 127). This observation led us to investigate whether or not such consistency remains valid over a significant amount of protein profiles. We ran the simulations over a thousand number of protein profiles and we identified that such observation holds. Therefore, denoting such *lag* by *offset* we postulate the following ***Offset-correction scheme*** in this context: *add one with the SCICA's outputs to obtain the original protein profiles* as shown in Fig. 1(d), (h), (l), (p).

As mentioned in Section II-C, the present study assumes the input signal is the frequency domain representation of a mixture of independent signals. Each of these signals will be an oscillatory signal whose frequency gives the position of the peaks that in turn represents the protein profile as shown in Fig. 1. This algorithm requires only one spectrum observation as the input as opposed to the conventional ICA and thereby capable of saving significant amount of experimental time and the proposed statistical offset correction scheme makes the extraction method robust. This method is based on statistical observation with a sample space of 830 samples. However, the mathematical basis to understand such offset occurring in this context forms part our future research.

IV. Conclusion

In this paper we proposed the application of SCICA, for the first time to the best of our knowledge, in the area of protein signal processing especially in extracting protein profiles from the mixed mass-spectra. The robustness of the algorithm is also ensured by our proposed offset correction scheme based on the statistical data analysis. We envisage the proposed concept will expedite the mass-spectra data analysis as well as will help discover robust bio-markers useful for different types of cancer detection. Our future research constitutes of proving this proposed concept on more number of simulated and real proteins and its on-field deployment.

References

[1] Mantini et al (2008)., Independent component analysis for the extraction of reliable protein signal profiles from MALDI-TOF mass spectra. Bioinformatics, Vol. 24, no. 1, pages 63–70.

[2] Gras,R. et al. (1999), Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. Electrophoresis, 20, 3535–3550.

[3] Coombes,K.R. et al. (2005) Improved peak detection and quantification of mass spectrometry data acquired from SELDI by denoising spectra with the undecimated discrete wavelet transform. Proteomics, 5, 4107–4117.

[4] Mantini,D. et al. (2007) LIMPIC: a computational method for the separation ofprotein signals from noise. BMC Bionformatics, 8, 101.

[5] Satten,G.A. et al. (2004) Standardization and denoising algorithms for mass spectra to classify whole-organism, 20(17):3128-3136.

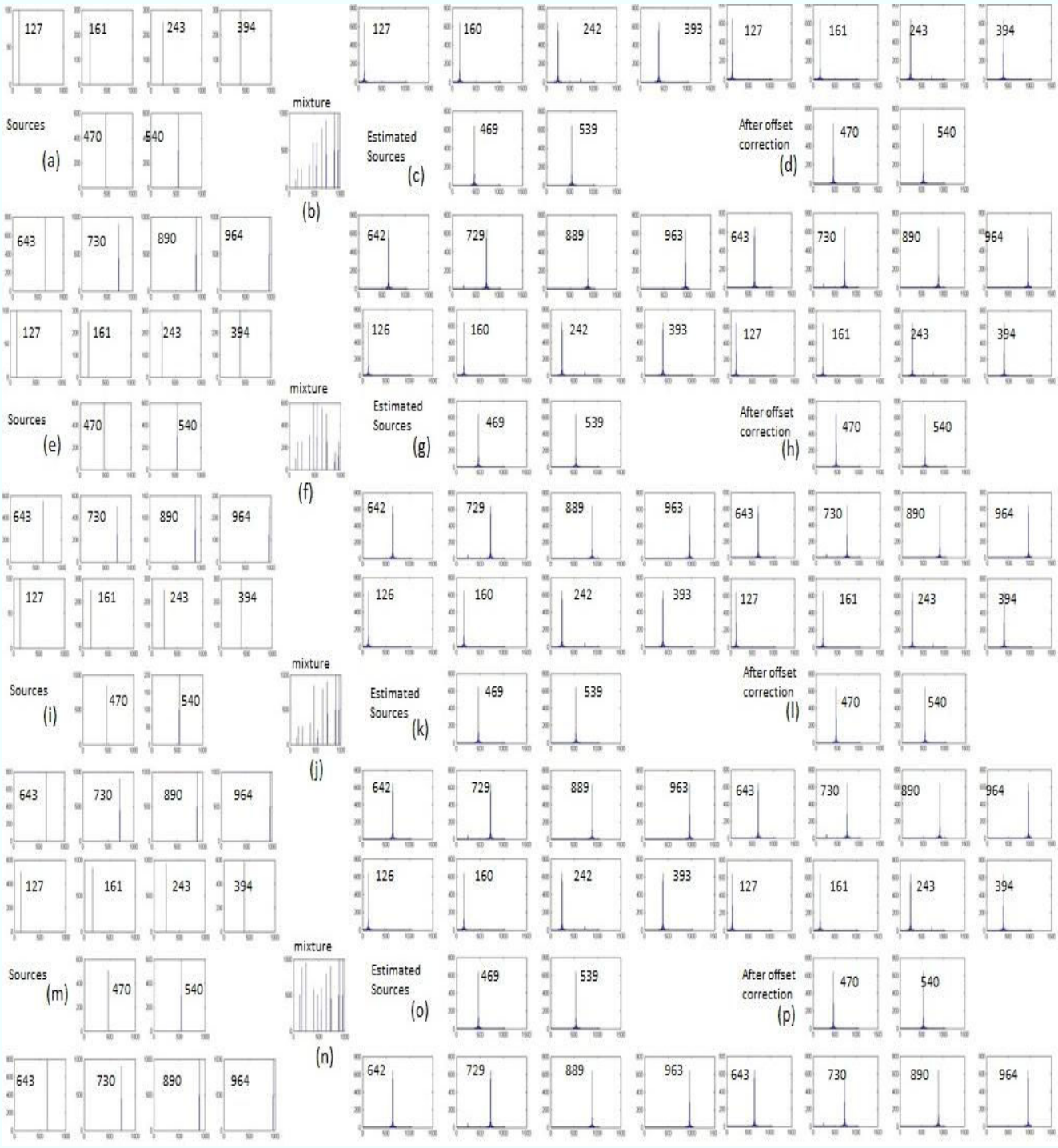

Fig. 1(a),(e),(i),(m): Original protein profiles for four different cases; (b),(f),(j),(n): simulated mass spectra (Mixed signal); (c),(g),(k),(o): SCICA's outputs; (d),(h),(l),(p): Extracted protein profiles after offset-correction.

[6] Yasui,Y. et al. (2003) An automated peak identification/calibration procedure for high-dimensional protein measures from mass spectrometers. J. Biomed. Biotechnol., 4, 242–248.

[7] M.E. Davies et. Al.(2004), A simple mixture model for sparse over-complete ICA, IEE Proc. VISP 151 (1) 35–43.

[[8] M. Girolami(2002), A variational method for learning sparse and over-complete representations, Neural Comput. 13 (112) 517–532.

[9] S Wang et. al., (2007), On the ICA of evoked potentials through single or few recording channel,. IEEE EMBS, France, August 23-26, 2007.

[10] Foley,J.P. (1987) Equations for chromatographic peak modeling and calculationof peak area. Anal. Chem., 59, 1984–1987.