The 2010 Signal Separation Evaluation Campaign (SiSEC2010): Biomedical Source Separation

Shoko Araki[1], Fabian Theis[2], Guido Nolte[3], Dominik Lutter[2], Alexey Ozerov[4], Vikrham Gowreesunker[5], Hiroshi Sawada[1], and Ngoc Q.K. Duong[4]

[1] NTT Communication Science Labs., NTT Corporation, Japan
[2] IBIS, Helmholtz Zentrum München, Germany
[3] Fraunhofer Institute FIRST IDA, Germany
[4] INRIA, Centre Inria Rennes - Bretagne Atlantique, France
[5] DSPS R&D Center, Texas Instruments Inc., USA

Abstract. We present an overview of the biomedical part of the 2010 community-based Signal Separation Evaluation Campaign (SiSEC2010), coordinated by the authors. In addition to the audio tasks which have been evaluated in the previous SiSEC, SiSEC2010 considered several biomedical tasks. Here, three biomedical datasets from molecular biology (gene expression profiles) and neuroscience (EEG) were contributed. This paper describes the biomedical datasets, tasks and evaluation criteria. This paper also reports the results of the biomedical part of SiSEC2010 achieved by participants.

1 Introduction

Large-scale evaluations are a key ingredient to scientific and technological maturation by revealing the effects of different system designs, promoting common test specifications and attracting the interest of industries and funding bodies. Recent evaluations of source separation systems include a series of the BCI (Brain Computer Interface) competitions [1, 2, 3].

The former community-based Signal Separation Evaluation Campaign (SiSEC 2008) [4] was designed based on the panel discussion at the 7th International Conference on Independent Component Analysis and Signal Separation (ICA 2007) which featured the Stereo Audio Source Separation Evaluation Campaign [5]. The general principles of the SiSEC aim to facilitate the entrance of researchers addressing different tasks and to enable detailed diagnosis of the submitted systems. The unique aspect of the SiSEC is that, SiSEC is not a competition but a scientific evaluation from which we can draw rigorous scientific conclusions.

The former SiSEC2008 attracted around 30 entrants, however, it consisted solely of audio datasets and tasks. Obviously, in addition to audio, there are many important areas where signal separation techniques are contributing to successful analyses. According to such a feedback to SiSEC2008, it was decided

V. Vigneron et al. (Eds.): LVA/ICA 2010, LNCS 6365, pp. 123–130, 2010.

that SiSEC2010 included not only the audio datasets, but also the biomedical datasets and tasks. For the first biomedical evaluation campaign of in SiSEC, we had three datasets: a set of gene expression profiles and two EEG datasets.

With the advent of high-throughput technologies in molecular biology, genome-wide measurements of gene transcript levels have become available. It has become clear that efficient computational tools are needed to successfully interpret the information buried in those large scale gene expression patterns. One commonly taken approach is to apply exploratory machine learning to detect similarities in multiple expression profiles in order to group genes into functional categories — for example, genes that are expressed to a greater or lesser extent in response to a drug or an existing disease. Although classically hierarchical clustering techniques are mostly used for these analyses, it has recently been shown [6, 7] that factorization techniques such as ICA can successfully separate meaningful categories by considering gene expression patterns as a superposition of independent expression modes, which are considered putative independent biological processes. Here we want to evaluate blind source separation problems applied to a set of microarrays in which we can quantify the expected cluster properties. We will see that inclusion of additional biological information as prior will be key to successful separations.

Most commonly, biomedical campaigns for electrophysiological data are formulated for BCI tasks, where an algorithm has to estimate e.g. the side of actual movement or imagined movement of left or right fingers of a subject from EEG data [1, 2, 3]. Here, the objective is clear and can be formulated using real EEG data only. The situation is more difficult in the case of decomposition of real data because the ground truth is not known and any simulation is prone to miss important aspects of real data. Here, two approaches are possible: a) one puts much effort to simulate *all* aspects of real EEG data [8], as imperfect as that may be, or b) one modifies real data as little as possible to construct some ground truth which can be tested for. We here take the second approach, specifically addressing the question whether independent sources can be distinguished from dependent ones.

This article describes the biomedical tasks, which are newly considered in the SiSEC2010. The traditional "audio source separation" task in SiSEC2010 is described in [9]. Section 2 describes the biomedical datasets, tasks, criteria. We summarize the results for each task in Section 3. In this paper, we focus on the general specifications and outcomes of the campaign and let readers refer to `http://sisec.wiki.irisa.fr/` for more detail.

2 Biomedical Separation Task Specifications

SiSEC2010 includes three biomedical tasks. The remaining part of this section provides the explanations about the datasets, tasks and evaluation criteria for each task. All materials, including data and reference codes, are available on the website at `http://sisec.wiki.irisa.fr/`.

2.1 Task 1: Source Identification in Microarray Data

Dataset. Since microarray technology has become one of the most popular approaches in the field of gene expression analysis, numerous statistical methods have been used to provide insights into the biological mechanisms of gene expression regulation [10, 11]. A common microarray dataset consists of multiple gene expression profiles. Each expression profile $\mathbf{x}_i$ mirrors the expression of N genes via measuring the level of the corresponding messenger RNA (mRNA) under a specific condition. In our case, mRNA was extracted from $i = 189$ invasive breast carcinomes [12] and measured using Affymetrix U133A Gene-chips. The Affymetrix raw data was normalized using the RMA algorithm [13] from the R Bioconductor package *simpleaffy*. Non-expressed genes were filtered out and Affymetrix probe sets were mapped to Gene Symbols. This resulted in a total of $N = 11815$ expressed genes.

The task. Cell signaling pathways play a major role in the formation of cancer. Understanding the biology of cell signaling helps to understand cancer and to develop new therapies. The regulation of these signaling pathways takes place on multiple layers, one of those is the regulation of gene expression or transcription. Single genes can take part in more than one pathway and the expression profiles can be regarded as linear superpositions of different signaling pathways or more generally biological processes. Using blind source separation (BSS) techniques, a linear mixture model can be decomposed to reconstruct source signals, which can be interpreted as these signaling pathways. A more detailed discussion of the linear factor model can be found in [6, 7].

The task is now to reconstruct these signaling pathways or parts of it from the microarray expression profiles using BSS techniques. Here, we approximate signaling pathways as simple gene lists. These pathway gene lists were taken from NETPATH (www.netpath.org).

Evaluation. To evaluate the quality of the reconstructed pathways, we tested for the significance of enriched genes that can be mapped to the pathways. For each source signal or estimated pathway we identify the number of genes that map to the distinct pathways and calculate p-values using Fisher's exact test. To correct for multiple testing we use the Benjamini-Hochberg procedure to estimate false positive rates (FDR). Now, after Benjamini-Hochberg correction a reconstructed pathway was declared as enriched if the p-value was below 0.05. We then count the number of all different significantly reconstructed pathways.

2.2 Task 2: Dependent Component Extraction

Dataset. In this set, data are constructed which are as close as possible to real EEG data with minor changes to ensure some ground truth which is, in principle, detectable. While independent sources are often a useful assumption recent research has focused on the analysis of brain connectivity. Therefore, the objective of the research is the detection of dependent sources.

The data were constructed as a superposition of $N = 19$ sources measured in as many sensors. Out of the 19 sources two were dependent and all others were mutually independent. The construction was done in the following way. We decomposed EEG data $\mathbf{x}(t) = (x_1(t), ..., x_N(t))^T$ from one subject using the TDSEP algorithm [14] as a Blind Source Separation in the standard way as

$$\mathbf{x}(t) = A\mathbf{s}(t) \tag{1}$$

with A being the mixing matrix and $\mathbf{s}(t)$ being the estimated independent source activities. Two of the 19 sources were selected as being dependent. For notational simplicity we denote the respective source indices as 1 and 2. The criterion for dependence was the fact that the imaginary part of the coherency showed a clear signal at around 10Hz. Generally, coherency between sources i and j at frequency f is defined as

$$C_{ij}(f) = \frac{< \hat{s}_i(f)\hat{s}_j^*(f) >}{\sqrt{< |\hat{s}_i(f)|^2 >< |\hat{s}_j(f)|^2 >}} \tag{2}$$

where $\hat{s}_i(f)$ is the short-time Fourier-transform of $s_i(t)$ for each trial (or segment) and $<>$ denotes averaging over all trials/segments. For these data we chose segments of 1 second duration giving a total of around 600 segments.

It can easily be shown that a significant non-vanishing imaginary part of $C_{ij}(f)$ cannot be explained by a mixture of independent sources [15]. The reason is that for mixtures of independent sources any coupling is caused by contributions of the same source which has a phase delay of 0 or π if the mixing coefficients have equal or opposite sign, respectively. In either case, the imaginary part vanishes.

For the present data the spatial patterns of the two ICA-components, i.e. the first two columns of the mixing matrix A, are shown in the upper panels of Fig.1. The patterns indicate clear signals from frontal and occipital parts of the brain, respectively. In the lower panels we show the power spectrum of the two ICA-components and the imaginary part of coherency between the components. Interestingly, the power at 10Hz is almost not visible in the frontal source, but the imaginary part of coherency still indicates a clear interaction between the sources at 10Hz.

The data $\mathbf{y}(t)$ were then constructed as

$$\mathbf{y}(t) = \sum_{i=1}^{2} \mathbf{a}_i s_i(t) + \sum_{i=3}^{N} \mathbf{a}_i \tilde{s}_i(t) \tag{3}$$

where $\mathbf{a}_i$ is the $ith.$ column of the mixing matrix A. The time series $\tilde{s}_i(t)$ were all taken from real data from different subjects. For each subject the data were decomposed using ICA and the $i.th$ original source $s_i(t)$ (for $i > 2$) was replaced by the $i.th$ source of the $i.th$ subject with ordering according to magnitude of the ICA-components.

The task. The task was to recover from $\mathbf{y}(t)$ the space spanned by the two columns $\mathbf{a}_1$ and $\mathbf{a}_2$. It was not the task to recover the two columns separately because for an interacting system the information given was not sufficient. It would have been necessary to make additional, e.g. spatial assumptions, on the nature of the sources to uniquely decompose the subspace into separate sources. Although in this special case, having distinct topographies at opposite sides of the scalp, such assumptions would have been both reasonable and easy to implement we preferred to avoid additional complications.

Evaluation. The geometrical relations between two subspaces can be defined in terms of the respective projectors on the respective subspace. If $\hat{A} = (\mathbf{a}_1, \mathbf{a}_2)$ then

$$P_A = \hat{A} \left(\hat{A}^T \hat{A} \right)^{-1} \hat{A}^T \tag{4}$$

is a projector onto the space spanned by the columns $\mathbf{a}_1$ and $\mathbf{a}_2$, i.e. P_A is a projector onto the true subspace. Similarly, let P_B be the projector on the estimated 2-dimensional subspace. We then calculate the eigenvalues of

$$D = P_A P_B P_A \tag{5}$$

Writing the eigenvalues in descending order for two-dimensional subspaces only the first two eigenvalues can be non-vanishing and all eigenvalues are in the interval $[0, 1]$. The subspaces are identical if and only if the second eigenvalue is equal to 1. For the data evaluation the value of this second eigenvalue was used to asses how accurately the true subspace was recovered.

2.3 Task 3: Artifact Removal in EEG Data

Dataset. This task contains two data sets: (1) 8-ch newborn EEG data, that are effectively only six channels due to the connection of electrodes, not containing any obvious artifact, and (2) an artificial data that represent artifacts of different kind: unleaded electrode, eye blinking. Some example figures can be found in `http://sisec.wiki.irisa.fr/`.

The data from active sleep of a newborn individual were sampled at 128 Hz and have total length of two minutes. There is a mutual dependence within the first four channels and also within the last four channels due to montage of electrodes, so that only six channels of eight are really independent.

The task. The data contains separately artifact and clean data, denoted a and x. The task is to blindly separate a and x from $y = a + x$. Participants have to estimate $\hat{x}$ that minimizes the norm $||\hat{x} - x||$. Of course, just y is available for the task.

Evaluation. We measure the Euclidean distance between the original and the reconstructed data.

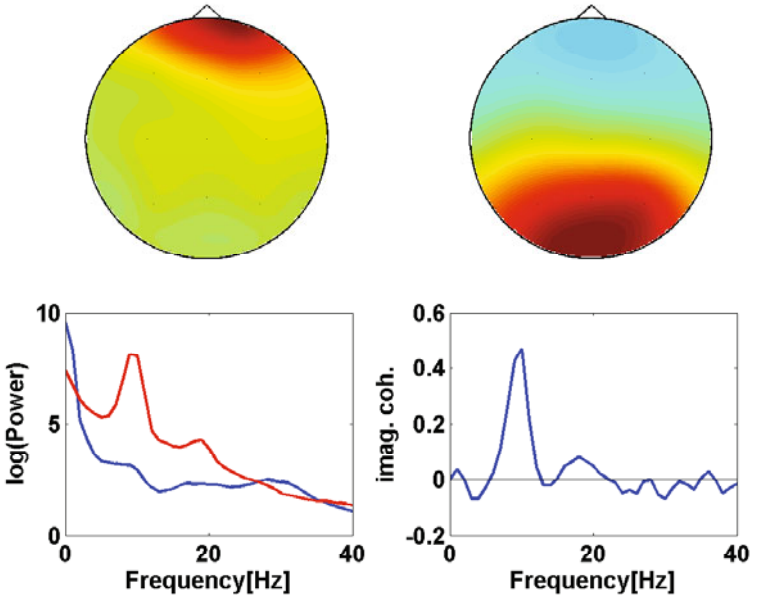

Fig. 1. Upper panels: topographies of selected ICA components chosen as interacting sources. Lower left panel: power spectrum of interacting sources. The blue line corresponds to the frontal source (upper left panel) and the red line, having a pronounced peak at around 10Hz, to the occipital source (upper right panel). Lower right panel: Imaginary part of coherency between these two sources.

3 Results

3.1 Task 1: Source Identification in Microarray Data

The task was processed by two independent groups. Chen et al. included prior information by using network component analysis (NCA) [16] to reconstruct the pathways. Blöchl et al. applied a matrix factorization method using prior knowledge encoded in a graph model (GraDe) [17], which was also presented at the LVA/ICA workshop 2010. Chen et al. obtained 58 cancer relevant pathways, out of which 8 have highly correlated expression profiles and were removed from the analysis to decrease redundancy. Blöchl et al. reconstructed 194 source profiles or pathways (one for each microarray measurement).

Both groups separated all 10 pathways with at least one source profile with a p-value below 0.05 (Fisher's exact test). Reduced to source profiles that specifically match to one pathway only, Blöchl et al. separated 7 of 10 pathways and Chen et al. 5 of 10. After FDR correction the number of enriched pathways reconstructed using GraDe reduced to 5 whereas whereas the NCA approach could not deliver any significantly enriched pathway. According to this task one can assess that the GraDe approach outperformed the NCA approach. We hypothesize that the better performance of the GraDe methods arises from the use

of a knowledge-based transcription factor network including pathway information (TRANSPATH [18]) in contrast to a transcription factor network as prior knowledge (TRANSFAC [19]) used in the NCA approach.

3.2 Task 2: Dependent Component Extraction

About task 2, we had one submission by Petr Tichavsky and Zbynel Koldovsky. The contributers applied ICA to the data and analyzed the dependence between the ICA components. Although they reported that the dependence was very weak they were able to pick the right components. With a second eigenvalue of 0.9998 the space was reconstructed almost perfectly.

3.3 Task 3: Artifact Removal in EEG Data

Unfortunately, task 3 did not receive any attempt to solve it. Perhaps, task 3 was too difficult, because a few channels were available and the dataset included a large number of artifacts.

4 Conclusion

This paper provided the specifications and results of the biomedical part of the SiSEC2010. Our dataset included not only the EEG data that were commonly addressed in previous biomedical campaigns, but also a microarray dataset. This time, we had three entrants. In order to obtain more participants, we have to reconsider how we can encourage to collaborate each other more efficiently. We invite all the willing participants to the continuous collaborative discussion on future of, both biomedical source separation and other separation tasks than biomedical and audio.

Acknowledgments. We would like to thank all the entrants, as well as P. Tichavsky for sharing his dataset and evaluation code.

References

1. Sajda, P., Gerson, A., Müller, K.R., Blankertz, B., Parra, L.: A data analysis competition to evaluate machine learning algorithms for use in brain-computer interfaces. IEEE Trans. Neural Sys. Rehab. Eng. 11(2), 184–185 (2003)
2. Blankertz, B., Müller, K.R., Curio, G., Vaughan, T.M., Schalk, G., Wolpaw, J.R., Schlögl, A., Neuper, C., Pfurtscheller, G., Hinterberger, T., Schröder, M., Birbaumer, N.: The BCI competition 2003: Progress and perspectives in detection and discrimination of EEG single trials. IEEE Trans. Biomed. Eng. 51(6), 1044–1051 (2004)
3. Blankertz, B., Müller, K.R., Krusienski, D., Schalk, G., Wolpaw, J.R., Schlögl, A., Pfurtscheller, G., Millán, J.R., Schröder, M., Birbaumer, N.: The BCI competition III: Validating alternative approachs to actual BCI problems. IEEE Trans. Neural Sys. Rehab. 14(2), 153–159 (2006)
4. Vincent, E., Araki, S., Bofill, P.: The 2008 signal separation evaluation campaign (SiSEC 2008. In: Adali, T., Jutten, C., Romano, J.M.T., Barros, A.K. (eds.) ICA 2009. LNCS, vol. 5441, pp. 734–741. Springer, Heidelberg (2009)

5. Vincent, E., Sawada, H., Bofill, P., Makino, S., Rosca, J.: First stereo audio source separation evaluation campaign: Data, algorithms and results. In: Davies, M.E., James, C.J., Abdallah, S.A., Plumbley, M.D. (eds.) ICA 2007. LNCS, vol. 4666, pp. 552–559. Springer, Heidelberg (2007)
6. Lutter, D., Ugocsai, P., Grandl, M., Orso, E., Theis, F., Lang, E., Schmitz, G.: Analyzing m-csf dependent monocyte/macrophage diffrentiation: expression modes and meta-modes derived from an independent component analysis. BMC Bioinformatics 9(100) (2008)
7. Teschendorff, A.E., Journ'ee, M., Absil, P.A., Sepulchre, R., Caldas, C.: Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Computational Biology 3(8) (2007)
8. Vidaurre, C., Nolte, G.: Generating synthetic EEG. In: Blankertz, B. (ed.) The BCI competition IV. LNCS, Springer, Heidelberg (to appear)
9. Araki, S., Ozerov, A., Gowreesunker, V., Sawada, H., Theis, F., Nolte, G., Lutter, D., Duong, N.Q.K.: The 2010 signal separation evaluation campaign (SiSEC2010): - audio source separation -. In: Vigneron, V. (ed.) LVA/ICA 2010. LNCS, vol. 6365, pp. 114–122. Springer, Heidelberg (2010)
10. Quackenbush, J.: Computational analysis of microarray data. Nature 2, 418–427 (2001)
11. Schachtner, R., Lutter, D., Knollmller, P., Tom, A.M., Theis, F.J., Schmitz, G., Stetter, M., Vilda, P.G., Lang, E.W.: Knowledge-based gene expression classification via matrix factorization. Bioinformatics 24, 1688–1697 (2008)
12. Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., Smeds, J., Nordgren, H., Farmer, P., Praz, V., Haibe-Kains, B., Desmedt, C., Larsimont, D., Cardoso, F., Peterse, H., Nuyten, D., Buyse, M., de Vijver, M.V., Bergh, J., Piccart, M., Delorenzi, M.: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. Journal of Natl. Cancer Inst. 98(4), 262–272 (2006)
13. Irizarry, R., Bolstad, B., Collin, F., Cope, L., Hobbs, B., Speed, T.: Summaries of affymetrix genechip probe level data. Nucleic Acid Research Journal 31(4) (2003)
14. Ziehe, A., Mueller, K.: TDSEP - an effient algorithm for blind separation using time structure. In: Proc. of ICANN 1998, pp. 675–680 (1998)
15. Nolte, G., Bai, U., Weathon, L., Mari, Z., Vorbach, S., Hallet, M.: Identifying true brain interaction from EEG data using the imaginary part of coherency. Clinical Neurophysiology (115) (2004)
16. Chen, W., Chang, C., Hung, Y.: Transcription factor activity estimation based on particle swarm optimization and fast network component analysis. IEEE EMBC 2010 (2010) (submitted)
17. Blöchl, F., Kowarsch, A., Theis, F.J.: Second-order source separation based on prior knowledge realized in a graph model. In: Vigneron, V. (ed.) LVA/ICA 2010. LNCS, vol. 6365, pp. 434–441. Springer, Heidelberg (2010)
18. Krull, M., Pistor, S., Voss, N., Kel, A., Reuter, I., Kronenberg, D., Michael, H., Schwarzer, K., Potapov, A., Choi, C., Kel-Margoulis, O., Wingender, E.: Transpath: an information resource for storing and visualizing signaling pathways and their pathological aberrations. Nucleic Acids Res. 34(Database issue), D546–D551 (2006)
19. Matys, V., Fricke, E., Geffers, R., Gssling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Mnch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S., Wingender, E.: Transfac: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 31(1), 374–378 (2003)