The 2011 Signal Separation Evaluation Campaign (SiSEC2011): - Biomedical data analysis -

Guido Nolte[1], Dominik Lutter[2], Andreas Ziehe[3], Francesco Nesta[4], Emmanuel Vincent[5], Zbynek Koldovsky[6], Alexis Benichoux[5] and Shoko Araki[7]

[1] Fraunhofer Institute FIRST, Germany
[2] IBIS, Helmholtz Zentrum München, Germany
[3] Technical University Berlin, Germany
[4] Fondazione Bruno Kessler - Irst, Center of Information Technology, Italy
[5] INRIA, Centre Inria Rennes - Bretagne Atlantique, France
[6] Technical University of Liberec, Czech Republic
[7] NTT Communication Science Labs., NTT Corporation, Japan

Abstract. This paper summarizes the bio part of the 2011 community based Signal Separation Evaluation Campaign (SiSEC2011). Two different data sets were given. In the first task, participants were asked to estimate the causal relations of underlying sources from simulated bivariate EEG data. In the second task, participants were asked to reconstruct signaling pathways or parts of it from the microarray expression profiles. The results for each task were evaluated using different objective performance criteria. We provide an overview of the biomedical datasets, tasks and criteria, and we report on the achieved results.

1 Introduction

The Signal Separation Evaluation Campaign (SiSEC) is a regular campaign focused on the evaluation of methods for signal separation. While its main focus is on separation of audio data, after the campaign in 2010 [1] this is now the second time that tasks on biomedical data analysis are proposed. This article describes the bio part of SiSEC 2011.

The standard application of ICA-algorithms in biomedical data analysis are EEG and MEG data. In contrast to signal separation in audio datasets, the respective mixing model is static. The algorithms to solve such a problem are well established and are applied routinely by many researches. It is our opinion that conceptually only minor technical details could be added to present day knowledge. Additionally, a formulation of an ICA challenge for EEG/MEG data is problematic because of two reasons: a) in contrast to audio data, for real EEG/MEG data the ground truth is almost never known, and b), existing ICA algorithms exploit different statistical properties, and the winning method for simulated data will then be the one for which, essentially by coincidence, the simulated statistical properties match the exploited ones.

We therefore decided to deviate from the 'standard' problem and to propose two different tasks. In the first task, source separation shall be applied to analyze gene expressions, and in the second we simulate EEG data, but the task is not to separate sources but to separate the effect of confounding noise in an estimate of causal relations.

Details of the tasks can be found at http://sisec.wiki.irisa.fr/ and following the link to 'biomedical data analysis'.

2 Estimating causal relations

2.1 Task

Noninvasive electrophysiological measurements like EEG/MEG measure to large extent unknown superpositions of very many sources. Any relation observed between channels is dominated by meaningless mixtures of mainly independent sources. The question is how to observe and properly interpret true interactions in the presence of such strong confounders. Since recently, a focus of research are the causal relations between groups of neurons. Many methods have been suggested to address this question for EEG or MEG data [2,3,4,5,6].

In this task contributors are requested to estimate the direction of interaction for simulated unidirectional bivariate dynamical systems. The difficulty is the presence of additive noise which is both non-white and spatially correlated.

The task is to estimate the direction of the interaction of the signal. A submitted result is a vector with 1000 numbers having the values 1, -1, or 0. Here, 1 means direction is from first to second sensor, -1 means direction is from second to first sensor, and 0 means 'I do not know'.

2.2 Dataset

The dataset consists of 1000 examples of bivariate data for 6000 time points. Each example is a superposition of a signal (of interest) and noise. The signal is constructed from a unidirectional bivariate AR-model of order 10 with (otherwise) random AR-parameters and uniformly distributed input. The noise is constructed of three independent sources, generated with 3 univariate AR-models with random parameters and uniformly distributed input, which were instantaneously mixed into the two sensors with a random mixing matrix. The relative strength of noise and signal was set randomly. The Matlab code used to generate the data was provided. Note, that the phrase 'simulated EEG data' is meant loosely. The simulation addresses the conceptual problems of EEG data, but e.g. the actual spectra can be quite different from real EEG data.

The data $\mathbf{z}(t)$ were generated as

$$\mathbf{z}(t) = (1-\gamma)\frac{\mathbf{x}(t)}{||X||} + \gamma\frac{B\mathbf{y}(t)}{||BY||} \tag{1}$$

where $\mathbf{x}$ is a unidirectional linear system and $\mathbf{y}$ are two independent noise sources which are mixed into channels by a random matrix B. The parameter γ was set

randomly between 0 and 1, $||\cdot||$ denotes Frobenius matrix norm, and X and Y denote the full data as a matrix, e.g. $X = (\mathbf{x}(1), \mathbf{x}(2), ..., \mathbf{x}(N))$ for N data points. The noise $\mathbf{y}(t)$ was generated with an AR(10)-model with diagonal but otherwise random parameters and uniformly distributed input, i.e.

$$y_i(t) = \sum_{p=1}^{10} A_i(p)(t-p) + \eta_i(t) \tag{2}$$

for $i = 1, 2, 3$. For each data set the parameters A_{ik} were selected randomly according to a Gaussian distribution with a standard deviation 0.25. Nonstationary, i.e. diverging, systems were excluded. If the standard is substantially larger, almost all systems are nonstationary. If it is chosen substantially smaller, the spectra are nearly white. The 'innovation' $\eta_i(t)$ was uniformly distributed in the range $[-.5, .5]$. This takes into account that some algorithms require non-Gaussian data or, especially, non-Gaussian innovations.

The signal $\mathbf{x}(t)$ was generated in the following way. If, e.g., the first channel was the sender, then $x_1(t)$ was generated with a random AR-model of order 10 in the same way as the noise term, and $x_2(t)$ was generated as

$$x_2(t) = \sum_p A_{22}(p)x_2(t-p) + A_{21}(p)x_1(t-p) + \epsilon_2(t) \tag{3}$$

where, again, $\epsilon_2(t)$ was uniformly distributed in the range $[-.5, .5]$. The construction for the other direction is analogous.

2.3 Evaluation criterion

For all examples either 1 or -1 is correct. The most important point here is the way it is counted: you get +1 point for each correct answer; you get -10 points for each wrong answer; and you get 0 points for each 0 in the result vector. With this counting confidence about the result is added into the evaluation. It is strongly recommended that for each example the evidence for a specific finding is assessed. To our knowledge, this causality challenge is the first time that such an evaluation scheme is proposed.

2.4 Results

We received a total of 5 submissions. Results are shown in table1Another submission arrived after the deadline and after announcement of the results and was not counted. All participants were among the list of people who were contacted personally and were encouraged to submit.

This kind of challenge is new within the SiSEC campaign and can therefore not be compared to previous challenges.

Submission	Total Points	Correct Detections	False Detection
S1	-2289	701	299
S2 [7]	252	352	10
S3 [8]	-357	773	113
S4 [9,10,11,12,13]	218	278	6
S5 [14,15]	-247	163	41

Table 1. Results of causality challenge. The total points can be calculated as the number of correct detections minus ten times the number of false detections.

3 Cancer pathway reconstruction

3.1 The task

Cellular signaling pathways are the key transducers from extracellular signals to cellular reaction. Dysfunction of signaling pathways is often involved in the formation of cancer [16]. Thus, 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, from extracellular receptors to intracellular transduction, ending with the transcriptional activation of target genes. 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. All gene expression levels are represented by an $M \times N$ data matrix $\mathbf{X} = [\mathbf{x}_i^\top \dots \mathbf{x}_M^\top]$ with each row-vector $\mathbf{x}_m^\top$ representing the gene expression levels off all N genes measured in one experiment, or microarray. Assuming a linear mixture model, each vector $\mathbf{x}_m^\top$ represents a mixture of K unknown source signals $\mathbf{s}_k^\top$, each representing a pathway related gene expression profile with the corresponding mixing coefficients represented as a column-vector $\mathbf{a}_m$. Thus, using blind source separation (BSS) techniques, the data-matrix $\mathbf{X}$ can be decomposed into $\mathbf{X} = \mathbf{AS}$, where $\mathbf{A}$ is the $M \times K$ mixing matrix and $\mathbf{S}$ the $K \times N$ matrix of source signals. These source signals can now be used as a basis to identify distinct signaling pathways in terms of cellular responses [17]. A more detailed discussion of the linear factor model can be found in [18,19].

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

3.2 Dataset

The microarray technology a method for mRNA profiling has become one of the most popular approaches in the field of gene expression analysis. Based on the complexity of gene expression profiles, a variety of statistical methods have been developed to provide insights into the biological mechanisms of gene expression regulation [20,21,22]. The dataset consists of the i gene expression profiles. Each expression profile $\mathbf{x}_i$ mirrors the expression of N genes via measuring the level of the corresponding mRNA under a specific condition. In our case, mRNA

was extracted from $i = 189$ invasive breast carcinomes [23] and measured using Affymetrix U133A Gene-chips. The Affymetrix raw data was normalized using the RMA algorithm [24] 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.

3.3 Evaluation

Evaluation of the reconstructed pathways was performed by testing for the significance of enriched genes that can be mapped to the distinct 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. Finally, the number of all different significantly reconstructed pathways were counted.

3.4 Results

There were no submissions.

4 Conclusion

In this paper we presented the specifications of the biomedical data analysis part of SiSEC2011 and summarized the performance obtained over all the submissions. Two different tasks of very different nature we given. The 'Cancer pathway reconstruction' received no submission which could be due to the fact that the mathematical details were unclear to people not familiar with the biology.

For the EEG/MEG data analysis it might appear natural that ICA challenges were proposed. However, the ICA model for these data is not convolutive, which, from an algorithmic viewpoint, is a much simpler case than acoustic data. For instantaneous mixtures the algorithms have become standard. Probably everything which could be said , apart from minor details, was said already, and such a challenge does not attract researchers working on the technical aspects.

It was therefore decided to propose a different kind of challenge, in which causal direction in the presence of noise were to be estimated and in which evidence had to assessed for a successful submission. The large variation across final scores that it is largely unclear how to optimally solve this problem. Although the data were, strictly speaking, nonlinear (i.e. non-Gaussian), the nonlinearity was small, and people working on nonlinear methods were effectively left out. For the future we intend to expand the simulations such that both linear and nonlinear methods can reasonably be applied.

Acknowledgements. The authors gratefully acknowledge financial support by EU, BMBF and DFG. We would like to thank our co-authors for their permission to use material from joint publications.

References

1. Araki S, Theis F, Nolte G, Lutter D, Ozerov A, Gowreesunker V, Sawada H and Duong N.Q.K. The 2010 Signal Separation Evaluation Campaign (SiSEC2010): Biomedical Source Separation. In: Proc. LVA/ICA. (2010) 123–130
2. Chen Y, Bressler SL, Knuth KH, Truccolo WA, Ding M. Stochastic modeling of neurobiological time series: power, coherence, Granger causality, and separation of evoked responses from ongoing activity. Chaos. 2006 Jun;16(2):026113.
3. Kaminski M, Ding M, Truccolo WA, Bressler SL. Evaluating causal relations in neural systems: granger causality, directed transfer function and statistical assessment of significance. Biol Cybern. 2001 Aug;85(2):145-57.
4. Baccala LA, Sameshima K. Partial directed coherence: a new concept in neural structure determination. Biol Cybern. 2001 Jun;84(6):463-74.
5. Schreiber T. Measuring information transfer. Phys Rev Lett. 2000 Jul 10;85(2):461-4.
6. Nolte G, Ziehe A, Nikulin VV, Schlögl A, Krämer N, Brismar T, Müller KR (2008) Robustly estimating the flow direction of information in complex physical systems. Phys Rev Lett. 100:234101.
7. S. Hu, G. Dai, Q. Dai, G. Worrell, and H. Liang, "Causality analysis of neural connectivity: critical examination of existing methods and advances of new methods," IEEE Transactions on Neural Networks (Regular Paper), vol. 22, no. 6, pp. 829-844, June, 2011.
8. Leistritz, L., W. Hesse, Arnold, M., Witte, H.: Development of interaction measures based on adaptive non-linear time series analysis of biomedical signals. Biomedizinische Technik 51(2006), 64-69.
9. Chavez, M. and Martinerie, J. and Le Van Quyen,M., "Statistical assessment of nonlinear causality: application to epileptic EEG signals," J. Neurosci Methods, vol. 124, no. 2, pp.113-128, 2003
10. Palus, M. and Komarek, V. and Hrncir, Z. and Sterbova, K., "Synchronization as adjustment of infomation rates: Detection from bivariate time series," Phys. Rev. E, vol. 63, pages. 046211, 2001.
11. Prichard, D. and Theiler, J.,"Generalized redundancies for time series analysis," Physica D, vol. 84, pp. 476–493, 1995.
12. Theiler,J. and Eubank,S. and Longtin,A. and Galdrikian,B. and Farmer,J.D, "Testing for Nonlinearity in Time Series: The Method of Surrogate Data," Physica D, vol. 58, pp. 77-94, 1992.
13. Vakorin,V.A. and Krakovska,O.A. and McIntosh?,A.R., "Confounding effects of indirect connections on causality estimation," Journal of Neuroscience Methods, vol. 184, no. 1, pp. 152–160, 2009.
14. R. Vicente, M. Wibral, M. Lindner, and G. Pipa. "Transfer entropy-a model-free measure of effective connectivity for the neurosciences," RID e-1566-2011. Journal of Computational Neuro- science, 30(1), pp. 45–67, February 2011.
15. M. Wibral, B. Rahm, M. Rieder, M. Lindner, R. Vicente, and J. Kaiser. "Transfer entropy in magnetoencephalographic data: Quantifying information flow in cortical and cerebellar networks", RID e-1566-2011. Progress In Biophysics & Molecular Biology, 105(1-2), pp. 80-97, March 2011.
16. Hoffman BD, Grashoff C, Schwartz MA. Dynamic molecular processes mediate cellular mechanotransduction. Nature. 2011 Jul 20;475(7356):316-23.
17. Lutter D, Langmann T, Ugocsai P, Moehle C, Seibold E, Splettstoesser WD, Gruber P, Lang EW, Schmitz G. Analyzing time-dependent microarray data using independent component analysis derived expression modes from human

macrophages infected with F. tularensis holartica. J Biomed Inform. 2009 Aug;42(4):605-11
18. Lutter, D., Ugocsai, P., Grandl, M., Orso, E., Theis, F., Lang, E., Schmitz, G.: Analyzing m-csf dependent monocyte/macrophage differentiation: expression modes and meta-modes derived from an independent component analysis. BMC Bioinformatics 9(100) (2008)
19. A.E.Teschendorff, M. Journ'ee, P.A., Sepulchre, R., Caldas, C.: Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Computational Biology 3(8) (2007)
20. Quackenbush J. Computational approaches to analysis of DNA microarray data. Yearb Med Inform. 2006:91-103.
21. Schachtner R, Lutter D, Knollmüller P, Tomé AM, Theis FJ, Schmitz G, Stetter M, Vilda PG, Lang EW. Knowledge-based gene expression classification via matrix factorization. Bioinformatics. 2008 Aug 1;24(15):1688-97.
22. Kowarsch A, Blöchl F, Bohl S, Saile M, Gretz N, Klingmüller U, Theis FJ. Knowledge-based matrix factorization temporally resolves the cellular responses to IL-6 stimulation. BMC Bioinformatics. 2010 Nov 30;11:585.
23. 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, Van de Vijver MJ, Bergh J, Piccart M, Delorenzi M. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. J Natl Cancer Inst. 2006 Feb 15;98(4):262-72.
24. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP. Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res. 2003 Feb 15;31(4):e15.