HHS Public Access

Author manuscript

Conf Proc IEEE Eng Med Biol Soc. Author manuscript; available in PMC 2019 January 30.

Published in final edited form as:

Conf Proc IEEE Eng Med Biol Soc. 2014 ; 2014: 6651–6654. doi:10.1109/EMBC.2014.6945153.

Three-way Parallel Independent Component Analysis for Imaging Genetics Using Multi-Objective Optimization

Alvaro Ulloa[1,2], **Jingyu Liu**[2], **Victor Vergara**[2], **Jiayu Chen**[2], **Vince Calhoun**[1,2], and **Marios Pattichis**[1]

[1]Department of Electrical and Computer Engineering, University of New Mexico, Albuquerque, NM 87106, USA, [alvarouc,vcalhoun,mpattichis]@unm.edu

[2]The Mind Research Network, Albuqueque, NM 87106, USA, [aulloa,jchen,jliu,vvergara,vcalhoun]@mrn.org

Abstract

In the biomedical field, current technology allows for the collection of multiple data modalities from the same subject. In consequence, there is an increasing interest for methods to analyze multi-modal data sets. Methods based on independent component analysis have proven to be effective in jointly analyzing multiple modalities, including brain imaging and genetic data. This paper describes a new algorithm, three-way parallel independent component analysis (3pICA), for jointly identifying genomic loci associated with brain function and structure. The proposed algorithm relies on the use of multi-objective optimization methods to identify correlations among the modalities and maximally independent sources within modality.

We test the robustness of the proposed approach by varying the effect size, cross-modality correlation, noise level, and dimensionality of the data. Simulation results suggest that 3p-ICA is robust to data with SNR levels from 0 to 10 dB and effect-sizes from 0 to 3, while presenting its best performance with high cross-modality correlations, and more than one subject per 1,000 variables.

In an experimental study with 112 human subjects, the method identified links between a genetic component (pointing to brain function and mental disorder associated genes, including PPP3CC, KCNQ5, and CYP7B1), a functional component related to signal decreases in the default mode network during the task, and a brain structure component indicating increases of gray matter in brain regions of the default mode region. Although such findings need further replication, the simulation and in-vivo results validate the three-way parallel ICA algorithm presented here as a useful tool in biomedical data decomposition applications.

I. INTRODUCTION

Biomedical studies tend to collect data from multiple modalities (such as brain images and genetic data) and from the same participants. This multimodal data is expected to provide extensive insights into underlying biological mechanisms. However, there are significant challenges associated with identifying latent variables that capture relationships across modalities while revealing intrinsic information of the underlying signals.

Some recent methods that succesfully incorporate relationships among multiple modalities are based on the independent component analysis (ICA) model [1], [2], [3]. In ICA based methods, observed data is decomposed into maximally independent sources and relationships across modalities are derived from the mixing matrices used to reconstruct observations form sources. In [1], the authors combine multi-modal observations assuming the same mixing matrix, thus the same number of sources, for all modalities. A limitation of this early approach is that it will not be applicable to modalities that exhibit different numbers of sources or dissimilar mixing matrices. In [2], the authors present a probabilistic approach based on a modular Bayesian framework. This method has two configurations, one forces a common modal map (sources) and the other shares the same mixing matrix among modalities. Similar to [1], both of the configurations impose strong constrains on how the information is shared among modalities. In [3], the restrictions of [1] were relaxed by allowing non-perfect correlations and different number of sources. However, the approach only allows the analysis of two modalities.

The current paper extends the technique proposed in [3] by analyzing three modalities as in [4], and formulating a new multi-objective optimization framework. The work presented in this manuscript includes:

- **Application to three modalities:** Beyond [3], the current method analyzes three modalities which allows to detect direct and indirect cross-modal relationships.
- **Multi-objective optimization approach:** We introduce a multi-objective optimization framework consisting of six objective functions: three entropies from source signals and three cross-correlations between modalities. Along the Pareto front, we search for solutions that maximize entropy and adjust certain weighting parameters to generate a solution that increases correlation as to achieve a balance between the objectives.
- **Method validation based on a simulation framework:** We validate the method through simulations. We assess dependency of the solution as a function of effect size (Cohens d measure), SNR (noise added to observations), dimensionality (number of variables and observations) and correlation strength among modalities.
- **Application to imaging genetics:** The proposed method is applied to a dataset consisting of single nucleotide polymorphism (SNP), structural and functional MRI modalities, collected from 112 healthy volunteers

The rest of this paper is organized as follows: in section II, we describe the proposed algorithm, the simulation framework, and the data used in the application; in section III, we present the results; and in section IV, we discuss our results and state our conclusions.

II. MATERIALS AND METHODS

A. Single Modality Independent Component Analysis

To define the independent component analysis (ICA) method for multiple modalities, let $X^{(i)}$ denote the observed data matrix for modality i with n subjects (rows) and $m^{(i)}$ variables

(columns). For each modality, we decompose the observed data matrix as $X^{(i)} = A^{(i)}S^{(i)}$ where $[A^{(i)}]_{n \times c^{(i)}}$ denotes the mixing matrix, $[S^{(i)}]_{c^{(i)} \times m^{(i)}}$ denotes the source matrix, and $c^{(i)}$ denotes the number of sources. The columns of A represent the subjects' loading patterns; i.e., how each component is weighted across subjects, whereas the rows of S indicate each component, the weighted pattern of variables.

Within each modality, the ICA model assumes independent stationary sources following a non-Gaussian probability density function. The infomax algorithm [5], that solves ICA for one modality, attempts to maximize the entropy of the estimated sources in S as given by

$$\max_{W^{(i)}} \left\{ H\left(W^{(i)}\right) \right\}$$

where, $H(\cdot)$ denotes the entropy function and $W^{(i)\dagger} = A^{(i)}$, the pseudo-inverse. The estimated sources are derived from the observation data matrix using $S^{(i)} = W^{(i)}X^{(i)}$.

B. Three-way parallel Independent Component Analysis Using Multi-Objective Optimization.

The three-way parallel ICA (3pICA) searches for maximally independent sources exploiting relationships across modalities which are assessed through correlation between loading matrix columns.

The 3pICA algorithm seeks to solve a multi-objective optimization problem that maximizes

$$\max_{(p,q,r),\, W^{(1)},\, W^{(2)},\, W^{(3)}} \left\{ \beta^T \cdot \left[H\left(W^{(1)}\right), H\left(W^{(2)}\right), H\left(W^{(3)}\right), \rho^{1,2}_{p,q}, \rho^{2,3}_{q,r}, \rho^{1,3}_{p,r} \right] \right\} \quad (1)$$

where β is a scalarization vector that balances entropy and correlation objectives; p, q, r refer to component indices matching columns from $A^{(1)}$, $A^{(2)}$ and $A^{(3)}$ for which the correlation needs to be maximized; $W^{(1)}$, $W^{(2)}$, $W^{(3)}$ refer to the unmixing matrices from each modality; and $\rho^{i,j}_{p,q}$ denotes the squared correlation between the p$^{\text{th}}$ column of $A^{(i)}$ and the q$^{\text{th}}$ column of $A^{(j)}$.

This multi-objective formulation requires a solution that balances the objectives against each other. We propose solving this problem by giving preference to the entropy objectives and penalizing cross-modality enhancements whenever they interfere with the maximal entropy search. In other words, the algorithm favors the Pareto solution that maximizes entropy.

An initial estimate is obtained at the first step of each iteration by using maximally correlated component triplet to select candidates for the (p, q, r) indices. We then search the directions that maximize the entropy objectives $H(W^{(i)})$, $i = 1, 2, 3$ and use (p, q, r) to compute correlation gradients that maximize the ρ-variables in (1). The algorithm is described in Algorithm 1.

For updating the unmixing matrices, we use the entropy natural gradient [5] of each modality:

$$\nabla W^{(i)} = I + \left(1 - 2Y^{(i)}\right)U^{(i)T}, \quad (2)$$

where $Y = \frac{1}{1+e^{-U}}$, $U = WX + W_0$, and W_0 is a bias term. The algorithm computes the update for each pair of correlations and adds them into $\nabla A^{(i)}$, where

$$\nabla \rho^{i,j} = \left(A_q^{(j)} - \mu_q^{(j)} - \frac{Cov\left(A_p^{(i)}, A_q^{(j)}\right)\left(A_p^{(i)} - \mu_p^{(i)}\right)}{\sigma^2\left(A_p^{(i)}\right)}\right) \quad (3)$$

and $\mu_p^{(i)}$ denotes the mean of the pth column of $A^{(i)}$. In the algorithm, we do not update these correlation measures if they fall below a certain threshold s to avoid over-emphasis of non-significant correlations among modalities.

In the case of a reduction of entropy due to the cross-correlation update, we provide a dynamic adaptation of weights, λ_i to de-emphasize the correlation objectives in (1). The algorithm terminates when there is no significant change in the unmixing matrix updates.

Input: $X^{(1)}$, $X^{(2)}$, $X^{(3)}$, and s

Output: $W^{(1)}$, $W^{(2)}$, and $W^{(3)}$

Initialization: $W^{(i)} \leftarrow I$, $i = 1, 2, 3$;

While $\left\{\left\|\Delta W^{(1)}\right\|_F, \left\|\Delta W^{(2)}\right\|_F, \left\|\Delta W^{(3)}\right\|_F\right\} > \in$ **do**

$A^{(1)} \leftarrow W^{(1)\dagger}$, $A^{(2)} \leftarrow W^{(2)\dagger}$, $A^{(3)} \leftarrow W^{(3)\dagger}$;

Solve $\left\{p, q, r\right\} \leftarrow \underset{\{p,q,r\}}{\text{argmax}} \left\{\rho_{p,q}^{1,2} + \rho_{p,r}^{1,3} + \rho_{q,r}^{2,3}\right\}$;

for $i = 1, 2, 3$, $j = 2, 3, 1$, $x =$ p, q, y, $y = q, r, p$ **do**

Compute $\nabla W^{(i)}$ using eq. (2);

if $\sqrt{\rho_{x,y}^{i,j}} > s$ **then**

Compute $\nabla\rho^{i,j}$ and $\nabla\rho^{j,i}$ using eq. (3)

else

$\nabla\rho^{i,j} = 0$ and $\nabla\rho^{j,i} = 0$

end

end

$\nabla A^{(1)} = \nabla \rho^{1,2} + \nabla \rho^{1,3}$, $\nabla A^{(2)} = \nabla \rho^{2,1} + \nabla \rho^{2,3}$;

$\nabla A^{(3)} = \nabla \rho^{3,1} + \nabla \rho^{3,2}$;

for $i = 1, 2, 3$ **do**

if $\left\|\Delta W^{(i)}\right\|_F > \epsilon$ **then**

$W^{(i)} \leftarrow ((W^{(i)} + \nabla W^{(i)})^{-1} + \lambda_i \nabla A^{(i)})^{-1}$

if *Entropy decreases* **then**

$\lambda_i \leftarrow 0.9\lambda_i$,

end

end

end

end

Algorithm 1: 3pICA optimization procedure. Refer to subsection II–B for definitions.

C. Simulation Framework

Based on the ICA model, we first designed a total of six random sources for each modality drawn from Laplacian ($\mu = 0$, $b = 1$), Uniform (−1; 1), Exponential ($\lambda = 1$), Gaussian ($\mu = 0$, $\sigma = 1$) and Student-t ($\mu = 0$, $df = 2$) distributions, and the last one drawn from a bimodal Gaussian distribution.

The latter was specially designed to emulate the effect size in a component as measured by the Cohen's d, which is reflected in the distance between bimodal distribution peaks (e.g., in brain imaging data, the smaller peak is the mean activation of brain regions responding to stimuli, while the large peak is the mean of brain background activation). Then, we generated three random mixing matrices drawn from a zero mean and unit variance multivariate Gaussian distribution. These matrices are further projected using a singular value decomposition method to enforce the designed cross-modality correlation. We set the bimodal source to be linked among modalities. Through multiplying the mixing matrix ($n \times c$) with the sources ($c \times m$), we obtained the observation data matrix ($n \times m$), for each modality. Finally, we added Gaussian noise with variance σ_η^2 to the data matrix.

We adjusted the effect size by tuning the mean of the smaller peak in the following probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}}\left(0.99e^{-x^2/2} + 0.01e^{-(x-d)^2/2}\right), \quad (4)$$

where d implies the Cohen's d effect size measure. From the noisy observation matrix $\hat{X}^{(i)} = X^{(i)} + \eta$, we measured its signal to noise ratio (SNR) based on σ_X^2 and σ_η^2. To simulate high dimensionality effects, we adjusted the ratio between the number of variables to the number of subjects and computed the base 10 logarithm of the ratio (log-ratio).

Overall, we simulated three data modalities with 6 components and default settings as in Table I. It is important to note that in the dimensionality test, we increased the number of variables from the default settings (see column 5 of Table I) to 100,000, in order to test a wider range of log-ratios. We then changed one parameter of interest at a time while fixing all others to the default value and evaluated the effect of each parameter on the algorithm performance.

D. Dataset based on sMRI, fMRI, and SNPs

A total of 112 right handed healthy subjects between the ages of 18 and 58 (31.88±10.86) with no history of traumatic brain injury or other illness were drawn from the multisite Mind Clinical Imaging Consortium (MCIC) schizophrenia study [6]. A total of 68 participants were male and 44 female.

The dataset consists of three modalities: sMRI, fMRI, and SNPs. A T1-weighted sMRI was acquired at each site using an oblique axial gradient echo sequence (More detail in [7]). The fMRI data were collected using a block sensori-motor response task, and DNA was extracted from blood samples. Genotyping for all participants was performed at the Mind Research Network using the Illumina Infinium HumanOmni1-Quad assay covering 1,140,419 SNP loci. After a standard pruning procedure, it resulted in 777,635 SNP loci. The number of SNPs was further reduced by removing SNPs not within 200 bps of 15,908 gene transcription sites extracted from annotation data in the UCSC genome database. After this additional data reduction 65,492 SNP loci remained and were used in the analysis. For additional details on fMRI and SNP collection refer to [8].

III. RESULTS

A. Simulation Results

We present the simulation results in Fig. 1. We measure component accuracy and link estimation error as summarized in the caption of Fig. 1. For each parameter setting, we conducted 3pICA and separate ICAs for each modality. Median accuracy (dots in Fig. 1) and uncertainty level, in the form of inter-quartile ranges (whiskers in Fig. 1), were calculated after repeating the analysis 20 times.

In Fig. 1a, the size effect performance test suggests good estimations for 3pICA, median accuracy above 0.95 and link estimation error below 0.05, across the whole range of size effects. The performance of ICA was comparable to 3pICA for effect sizes higher than 3. In Fig. 1b, a cross-correlation above 0.3 was needed for 3pICA to yield better results than ICA. ICA exhibited invariance in its performance with this parameter. The flat trend reported in Fig. 1c suggested noise invariant properties for both methods. Finally, 3pICA required a dimensionality lower than 2.25 (at least one subject per 177 variables) to yield almost

perfect results, but performed above ICA for dimensionality lower than 3 (at least one subject per 1000 variables). A significant drop on performance was observed for high dimensionality in the simulated data.

B. Experimental Results on fMRI, sMRI, and SNP

We applied 3pICA setting 7 components for functional MRI, 15 from structural MRI and 55 for SNP data and detected an average correlation of 0.4 for the resulting component triplet. The spatial contents of functional and structural components intersect revealing parts of the default mode network (frontal gyrus, anterior cingulate, precuneus, and cingulate). The genetic component highlighted genes including PPP3CC, KCNQ5, and CYP7B1 which are directly associated with brain function or mental disorders. PPP3CC is involved in the downstream regulation of dopaminergic signal transduction, KCNQ5 is a member of the KCNQ potassium channel gene family that is differentially expressed in subregions of the brain, and CYP7B1 is involved in neurosteroid metabolism,[9, chapter 18].

IV. DISCUSSION AND CONCLUDING REMARKS

As we expected, 3pICA used information from all modalities to provide a solution that reveal connections among them. The simulation results suggested that 3pICA outperforms ICA in most scenarios or otherwise provides comparable performance. Results from the imaging genetics application example yielded reasonable results.

Performed simulations provided understanding of the algorithms behavior under the tested parameters. The effect size invariant property of 3pICA suggests that the estimation of one modality's sources can benefit from shared information contained on the other two modalities. Results indicate that shared information helped 3pICA overcoming the impact of effect size on the data better than regular ICA. The step-like behavior of 3pICA accuracy in Fig. 1b as a function of the cross-modality correlation is a clear indicator of the threshold s set to avoid overemphasis of correlation. Given 200 subjects we set a conservative threshold of 0.2 to attain a significance level of $\frac{0.01}{3}$. This behavior occurs since 3pICA behave as three separate ICAs below the threshold s. To investigate the performance boundary on noise robustness, we further tested negative SNRs and observed a significant drop in performance for both methods below −15 dB. However, the researcher should question data acquisition or pre-processing pipeline when the dataset presents such high levels of noise. Regarding the dimensionality test, the multi-modal method was affected by the reduced sample size because it has a direct impact on the correlation inference and likely affects the estimation accuracy. In case of log-ratios higher than 3, we recommend to consider variable pruning prior to 3pICA.

Finally, we used a real dataset application as a proof of concept, where results obtained from healthy subjects suggest a connection between gray matter concentrations and the subject's ability to focus as indicated by a suppression of the default mode network. The 3pICA uncovered additional association with a genetic component. Within this component, SNPs located at relevant genes exhibited noticeably larger weights and present expressions through the brain.

In summary, we comprehensively evaluated 3pICA under various scenarios. Simulation results suggested that the novel multi-modal method was insensitive to effect sizes from 0 to 3 and SNRs from 0 to 10 dB, but mainly affected by the dimensionality of the problem, performing best when the dataset holds at least one subject per 1,000 variables. The application results confirm that relevant information can be extracted by jointly analyzing three modalities.

Acknowledgments

This work was supported by NIH grants R33DA027626

REFERENCES

[1]. Calhoun V, Adali T, and Liu J, “A feature-based approach to combine functional mri, structural mri and eeg brain imaging data,” in Engineering in Medicine and Biology Society, 2006. EMBS ‘06. 28th Annual International Conference of the IEEE, 8 2006, pp. 3672–3675.

[2]. Groves AR, Beckmann CF, Smith SM, and Woolrich MW, “Linked independent component analysis for multimodal data fusion,” Neuroimage, vol. 54, no. 3, pp. 2198–2217, 2011. [PubMed: 20932919]

[3]. Liu J, Pearlson G, Windemuth A, Ruano G, Perrone-Bizzozero NI, and Calhoun V, “Combining fmri and snp data to investigate connections between brain function and genetics using parallel ica,” Human brain mapping, vol. 30, no. 1, pp. 241–255, 2009. [PubMed: 18072279]

[4]. Vergara VM, Ulloa A, Calhoun VD, Boutte D, Chen J, and Liu J, “A three-way parallel ica approach to analyze links among genetics, brain structure and brain function,” NeuroImage, pp. –, 2014.

[5]. Bell AJ and Sejnowski TJ, “An information-maximization approach to blind separation and blind deconvolution,” Neural computation, vol. 7, no. 6, pp. 1129–1159, 1995. [PubMed: 7584893]

[6]. Gollub RL, Shoemaker JM, King MD, White T, Ehrlich S, Sponheim SR, Clark VP, Turner JA, Mueller BA, Magnotta V et al., “The mcic collection: a shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia,” Neuroinformatics, vol. 11, no. 3, pp. 367–388, 2013. [PubMed: 23760817]

[7]. Segall JM, Turner JA, van Erp TG, White T, Bockholt HJ, Gollub RL, Ho BC, Magnotta V, Jung RE, McCarley RW et al., “Voxel-based morphometric multisite collaborative study on schizophrenia,” Schizophrenia bulletin, vol. 35, no. 1, pp. 82–95, 2009. [PubMed: 18997157]

[8]. Chen J, Calhoun VD, Pearlson GD, Ehrlich S, Turner JA, Ho B-C, Wassink TH, Michael AM, and Liu J, “Multifaceted genomic risk for brain function in schizophrenia,” NeuroImage, vol. 61, no. 4, pp. 866–875, 2012. [PubMed: 22440650]

[9]. McEntyre J and Ostell J, “The NCBI handbook,” 2002.

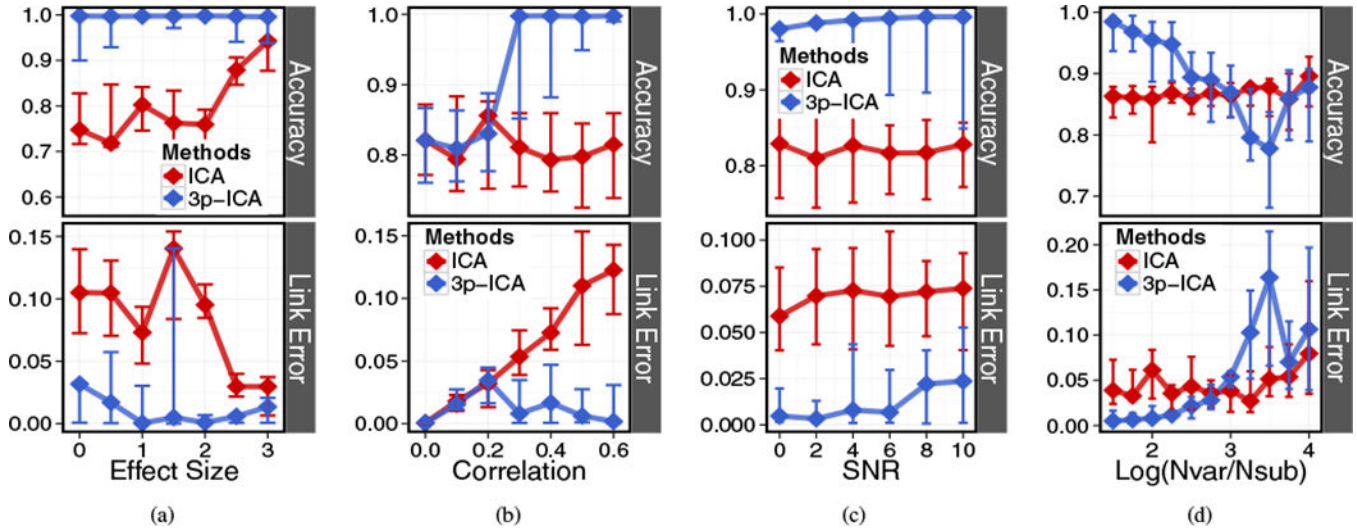

Fig. 1:

Simulation results from varying (a) effect size, (b) cross modality correlation, (c) SNR and (d) dimensionality. The upper plots show the accuracy of component estimation measured as the correlation between estimated and ground truth components. The lower plots show the absolute value of the difference between the estimated and designed correlations. The dots represent the median accuracy across 20 repetitions, and the whiskers indicate the inter-quartile range.

TABLE I:

Simulation settings.

Variable	Measure	Range	Step	Default
Effect Size	Cohen's d as in eq. (4)	[0, 3]	0.5	2
Correlation	$\rho_{p,q}^{1,2} = \rho_{q,r}^{2,3} = \rho_{p,r}^{1,3}$	[0, 0.6]	0.1	0.4
Noise level	$\mathrm{SNR}_{\mathrm{dB}}$	[0, 10]	2	10
Dimensionality	$\log_{10} \frac{\#\ Variables}{\#\ Subjects}$	[1.5, 4]	0.25	1.68