


APPROVAL SHEET

Title of Dissertation: Blind Source Separation for Multimodal Fusion of Medical Imaging Data

Name of Candidate: Yuri Levin-Schwartz
Doctor of Philosophy, 2017

Dissertation and Abstract Approved:


Dr. Tülay Adalı
Distinguished Professor
Department of Computer Science and
Electrical Engineering

Date Approved: 4/28/2017

ABSTRACT

Title of Dissertation: Blind Source Separation for Multimodal Fusion of Medical Imaging Data

Yuri Levin-Schwartz, Doctor of Philosophy, 2017

Dissertation directed by: Dr. Tülay Adalı, Professor
Department of Computer Science and
Electrical Engineering

Due to the ability of different sensors to provide complementary views of complicated systems, the collection of data from multiple sources has become common, particularly in neurological studies. Therefore, full utilization of the common information forms the fundamental goal of performing a joint analysis on this data. However, since little is known about the relationships among the datasets, it is important to minimize the underlying assumptions placed on the data. Because of this fact and their ability to treat separate datasets in a fully symmetric manner, multivariate data-driven methods are the main choice for the fusion of multiple sets of neurological data. However, different methods rely on different generative models, meaning that many considerations must be taken into account before an individual method can be applied to a new problem. This motivates an investigation of how the assumptions of different methods manifest in the resulting decomposition, the development of techniques to assess the contribution of different datasets to the analysis, and the development of a way to unambiguously assess the relative performance of different methods on real data.

In this dissertation, we approach these issues from multiple directions. We introduce a technique called principal component analysis and canonical correlation analysis (PCA-CCA) to determine the similarity or links between different neuroimaging datasets. Through both simulations and application to brain imaging data, namely, functional magnetic resonance imaging (fMRI) data, structural MRI (sMRI) data, and electroencephalo-

gram (EEG) data from patients with schizophrenia and healthy controls, we show the desirable performance of the proposed technique. We also use this unique set of imaging data in order to test the robustness of different data-driven fusion methods to deviations from their assumptions and to assess the effects of bringing each dataset into the analysis. We propose a classification rate-based procedure to quantify the performance of different fusion methods on real fMRI data extracted during the performance of different tasks and demonstrate how this method can be used to determine the “value added” by each dataset to the analysis. Finally, we introduce a novel visualization technique to highlight the changes in the brain regions that discriminate between patients with schizophrenia and healthy controls through the use of different fusion techniques. We find that the methods that we develop in this dissertation provide a useful framework for investigating the interactions of different datasets within a fusion analysis.

**Blind Source Separation for Multimodal Fusion of
Medical Imaging Data**

by
Yuri Levin-Schwartz

Dissertation submitted to the Faculty of the Graduate School
of the University of Maryland in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2017

In memory of my grandmothers Vivian Levin and Gertrude Schwartz.

ACKNOWLEDGMENTS

Though attempting to express my gratitude fully to all those who have played a role in making this dissertation a reality is a fool's errand, I will press forward nonetheless.

First, I must thank my advisor Professor Tülay Adalı for her guidance, enthusiasm, and patience throughout my doctoral work. Her emphasis on research as a collaborative activity has created a wonderful environment where questions and discussion are always encouraged. She has helped me to improve every aspect of my research immensely, from writing clearly and concisely to understanding new information quickly and thinking critically. I could not have asked for a better advisor.

I would like to thank the current and former members of the MLSP lab: Dr. Sai Ma, Dr. Wei Du, Dr. Gengshen Fu, Mohammad Abu Bak Siddique Akhonda, Zois Boukouvalas, Suchita Bhinge, Darren Emge, Qunfang Long, Rami Mowakeaa, and Xiaowei Song, for the countless hours of discussion, often long after the time that a reasonable person would go home.

I would also like to express my gratitude the members of my committee: Professor Vince Calhoun, Professor Joel Morris, Professor Seung-Jun Kim, and Professor Shari Waldstein for their insightful comments.

I owe my deepest thanks to my family. My parents have always been a source of encouragement and continue to inspire me. My sister, whose boundless enthusiasm and energy never fails to make me smile. I must also thank Robbin and Gene Warner for providing a "home away from home;" I cannot express my gratitude to them.

TABLE OF CONTENTS

DEDICATION	ii
ACKNOWLEDGMENTS	iii
LIST OF FIGURES	viii
LIST OF TABLES	.xviii
List of Abbreviations	xxi
Chapter 1 INTRODUCTION	1
1.1 Motivation	1
1.2 Contributions	3
1.3 Overview of Dissertation	4
Chapter 2 BACKGROUND	7
2.1 Neuroimaging Modalities	7
2.1.1 Feature Extraction	8
2.1.2 FMRI	9
2.1.3 EEG	12
2.1.4 SMRI	12

2.2	Dimension-Reduction and Order Selection	13
2.3	BSS Techniques for Data Fusion	15
2.3.1	JICA	16
2.3.2	CCA, MCCA, and IVA	20
2.3.3	TIVA	23
2.4	Minimum Spanning Tree (MST)	26
2.5	Summary	27
Chapter 3	ORDER-SELECTION FOR MULTIMODAL DATA	28
3.1	Introduction	29
3.2	Traditional and Sample-poor Hypothesis Test	31
3.2.1	Determining the Probability of False Alarm using ITC	33
3.3	Simulations	34
3.3.1	Variable Per-Component SNR Simulation Setup	34
3.3.2	Variable Per-Component SNR Simulation Results	36
3.3.3	Fixed Per-Component SNR Simulation Setup	37
3.3.4	Fixed Per-Component SNR Simulation Results	38
3.4	Fusion Results Using Real Multimodal Data	38
3.5	Discussion	41
3.6	Summary	44
Chapter 4	FUSION OF THREE MULTIMODAL NEUROIMAGING DATASETS	45
4.1	Introduction	46
4.2	Multimodal Data	47
4.3	Simulation	47

4.3.1	Simulation Setup	47
4.3.2	Simulation Results and Discussion	49
4.4	Multimodal Fusion of FMRI, SMRI, and ERP	51
4.4.1	JICA Results and Discussion	51
4.4.2	TCCA/TMCCA/TIVA Results and Discussion	53
4.5	Summary	57
Chapter 5	QUANTIFYING THE INTERACTION AND CONTRIBUTION OF MULTIPLE DATASETS IN FUSION	58
5.1	Introduction	59
5.2	Classification Procedure	60
5.2.1	Number of Monte-Carlo Resamplings	63
5.2.2	Order-Selection	64
5.3	Results	65
5.3.1	Stability of Conclusions Using Other Classification Techniques	68
5.3.2	Exploring the Effects of Fusion	70
5.4	Summary	73
Chapter 6	VISUALIZING THE CHANGES IN BIOMARKERS FROM FUSION	79
6.1	Motivation for and Generation of a GDM	80
6.1.1	Using GDMs to Summarize Discriminatory Power of a Decomposition	81
6.2	Visualizing Fusion	83
6.3	Summary	85

Chapter 7	CONCLUSIONS AND FUTURE WORK	86
7.1	Conclusions	86
7.2	Future Work	89
7.2.1	IVA for Determining the Connections between Multiple Datasets .	89
7.2.2	ICA+CCA	90
7.2.3	TIVA for Calcium Imaging Data	92
REFERENCES	93

LIST OF FIGURES

2.1	Extraction of features for each subject from multimodal neuroimaging data. Note the differences in the dimensionality of the original data and how a common <i>subjects</i> dimension is established through the extraction of features from each subject.	9
2.2	Generative model for ICA. Note how the m th column of the mixing matrix, \mathbf{a}_m , represents the relative weights of the m th source, \mathbf{s}_m , for each corresponding subject.	18
2.3	Generative model for multimodal fusion using tCCA. Note that this model is obtained after spatial dimension-reduction through PCA, <i>i.e.</i> , $\tilde{\mathbf{X}}^{[1]}$ and $\tilde{\mathbf{X}}^{[2]}$ are both $C \times M$ matrices.	24
3.1	Probability of correctly detecting the number of common components versus the dimension of the datasets, V_1 and V_2 , when the latent sources are (a) Gaussian, (b) uniformly, and (c) Laplacian distributed. Note that for these simulations, $V_1 = V_2$	36
3.2	Probability of correctly detecting the number of common components versus the dimension of the datasets, V_1 and V_2 , when the latent sources are (a) Gaussian, (b) uniformly, and (c) Laplacian distributed. Note that for these simulations, $V_1 = V_2$	37
3.3	Meaningful and statistically significant components generated using tCCA with pairwise combinations of the three modalities. The uncorrected significance (p -values) for the components are (a) 0.027 for the fMRI-sMRI component, (b) 0.003 for the fMRI-ERP component, and (c) 0.018 for the sMRI-ERP component. All spatial maps are Z-maps threshold at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients.	39

3.4	Meaningful and statistically significant components generated using MCCA with the GENVAR cost. The uncorrected significance (p -values) for the components are (a) 0.014 for the first component and (b) 0.018 for the second component. All spatial maps are Z -maps thresholded at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients.	42
4.1	Estimation performance for the common component with two datasets as (a) the height of the step in $\mathbf{a}_1^{[1]}$ decreases (b) $\mathbf{n}_1^{[1]}$ increases, and (c) the number of subjects increases. . . .	49
4.2	Estimation performance for the common component with three datasets as the height of the step decreases in terms of (a) t -statistic for the most significant component, (b) and (c) correlation of the most significant component to the original source.	50
4.3	Significant and biologically meaningful components generated for fusion of all three datasets using jICA using the EBM algorithm. The uncorrected significance values are (a) 0.0042 and (b) 0.0097. All spatial maps are Z -maps thresholded at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients. .	52
4.4	Significant and biologically meaningful components generated using jICA using the EBM algorithm for combinations of two modalities. The uncorrected significance values for the components are (a) 0.0031 for the first component and (b) 0.0147 for the second component from the fusion of fMRI and ERP, as well as (c) 0.0044 for the component from the fusion of sMRI and ERP. All spatial maps are Z -maps thresholded at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients. .	53

4.5	Significant and biologically meaningful components generated from the fusion of all three datasets using tMCCA using the MAXVAR and GENVAR cost functions as well as tIVA-G. The uncorrected significance values for the components estimated using MAXVAR are 0.0140. The uncorrected significance values for the components estimated using GENVAR 0.0134 for the top component and 0.0187 for the second one. For tIVA-G, we have a single component that shows significant group difference with 0.0265. All spatial maps are Z-maps thresholded at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients.	54
4.6	Significant and biologically meaningful components generated using tIVA-G and tCCA with two modalities. The uncorrected significance for components are 0.0130 for the component estimated using tIVA-G and 0.0130 for the component estimated using tCCA. All spatial maps are Z-maps thresholded at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients.	56

5.1	Classification process for a single feature dataset. For the case where multiple datasets are analyzed, the ICA step is replaced with either jICA or IVA, performed on the concatenated feature datasets or the collection of feature datasets, respectively. The procedure is as follows: (a) the data is split into a training set, $\mathbf{X}_{\text{Train}}$, and a test set, \mathbf{X}_{Test} . (b) the training dataset is dimension reduced using PCA, ICA is run, and the discriminatory components, $\check{\mathbf{S}}_{\text{Train}}$, and corresponding subject covariations, $\check{\mathbf{A}}_{\text{Train}}$, are selected. (c) in the final stage, $\check{\mathbf{A}}_{\text{Train}}$ is used to train the classifier, $\check{\mathbf{S}}_{\text{Train}}$ is regressed onto \mathbf{X}_{Test} producing $\check{\mathbf{A}}_{\text{Test}}$, and $\check{\mathbf{A}}_{\text{Test}}$ is used to test the classifier. This process is repeated N times and the mean classification rate is evaluated. Note that for IVA and well as the data integration technique, there are dataset-specific subject covariations $\check{\mathbf{A}}_{\text{Train}}^{[k]}$ and $\check{\mathbf{A}}_{\text{Test}}^{[k]}$; however, for IVA the subject covariations are derived by fusing information across all datasets whereas for the data integration technique the subject covariations are extracted from each dataset individually.	61
5.2	Average classification rate using KSVM and IVA on the combination of AOD and SIRP datasets as a function of Monte-Carlo resamplings.	63
5.3	Average classification results using ICA-EBM and KSVM on the AOD dataset for different orders.	64
5.4	Average classification results using KSVM for individual datasets and combinations of datasets using either data fusion, with jICA and IVA-GL, or data integration using combined ICAs. The first three columns from the left refer to the case where only one dataset is analyzed. The fourth, fifth, and sixth columns from the left refer to combinations of two datasets. The rightmost point shows the classification performance when all three task datasets are jointly analyzed. Note that error bars are omitted for clarity, since the largest value of the standard error is 0.0035.	66

5.5	Average classification results using KNN for individual datasets and combinations of datasets using either data fusion, with jICA and IVA-GL, or data integration using combined ICAs. The first three columns from the left refer to the case where only one dataset is analyzed. The fourth, fifth, and sixth columns from the left refer to combinations of two datasets. The rightmost point shows the classification performance when all three task datasets are jointly analyzed. Note that error bars are omitted for clarity, since the largest value of the standard error is 0.0034.	68
5.6	Average classification results using FDA for individual datasets and combinations of datasets using either data fusion, with jICA and IVA-GL, or data integration using combined ICAs. The first three columns from the left refer to the case where only one dataset is analyzed. The fourth, fifth, and sixth columns from the left refer to combinations of two datasets. The rightmost point shows the classification performance when all three task datasets are jointly analyzed. Note that error bars are omitted for clarity, since the largest value of the standard error is 0.0037.	70

5.7 Statistically significant components for the combination of the AOD and SIRP datasets.

(a) the significant components obtained through the use of data fusion, using IVA-GL, are shown in the first two rows. The third and fourth rows contain those significant components obtained through the use of data integration, using ICA-EBM, that have a correlation above 0.5 with the components obtained using IVA-GL. The aligned components are in the same column. Those components obtained using ICA-EBM that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (b) and (c) for the AOD and SIRP datasets, respectively. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients. The p -values for each component are located above the corresponding spatial map and those that remain significant after a Bonferroni correction are displayed in green. All spatial maps are Z -maps thresholded at $Z=2.7$ 74

5.8	Statistically significant components for the combination of the AOD and SM datasets. (a) the significant components obtained through the use of data fusion, using IVA-GL, are shown in the first two rows. The third and fourth rows contain those significant components obtained through the use of data integration, using ICA-EBM, that have a correlation above 0.5 with the components obtained using IVA-GL. The aligned components are in the same column. Those components obtained using ICA-EBM that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (b) and (c) for the AOD and SM datasets, respectively. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients. The p -values for each component are located above the corresponding spatial map and those that remain significant after a Bonferroni correction are displayed in green. All spatial maps are Z -maps thresholded at $Z=2.7$	75
-----	---	----

5.9 Statistically significant components for the combination of the AOD and SIRP datasets.

(a) the significant components obtained through the use of data fusion, using IVA-GL, are shown in the first two rows. The third through sixth rows contain those significant components obtained through the use of jICA that have a correlation above 0.5 with the components obtained using IVA-GL. The aligned components are in the same column. Note that due to the assumption by jICA of a common mixing matrix for all datasets, multiple spatial maps are considered a single component with one corresponding p -value. The seventh and eight rows contain the significant components obtained through the use of data integration, using ICA-EBM, that have a correlation above 0.5 with the components obtained using IVA-GL. The components obtained using jICA that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (b). Additionally, those components obtained using ICA-EBM that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (c) and (d) for the AOD and SIRP datasets, respectively. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients. The p -values for each component are located above the corresponding spatial map and those that remain significant after a Bonferroni correction are displayed in green. All spatial maps are Z-maps thresholded at $Z=2.7$ 76

5.10 Statistically significant components for the combination of the AOD and SM datasets.

(a) the significant components obtained through the use of data fusion, using IVA-GL, are shown in the first two rows. The third through sixth rows contain those significant components obtained through the use of jICA that have a correlation above 0.5 with the components obtained using IVA-GL. The aligned components are in the same column. Note that due to the assumption by jICA of a common mixing matrix for all datasets, multiple spatial maps are considered a single component with one corresponding p -value. The seventh and eight rows contain the significant components obtained through the use of data integration, using ICA-EBM, that have a correlation above 0.5 with the components obtained using IVA-GL. The components obtained using jICA that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (b). Additionally, those components obtained using ICA-EBM that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (c) and (d) for the AOD and SM datasets, respectively. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients. The p -values for each component are located above the corresponding spatial map and those that remain significant after a Bonferroni correction are displayed in green. All spatial maps are Z-maps thresholded at $Z=2.7$ 77

6.1 Comparison of the t -statistics from the GDMs compared with the t -statistics of the original biomarkers found using (a) ICA-EBM and (b) IVA-GL for each task. The t -statistics of the GDMs are shown in red, while the original biomarkers are shown in blue. 81

6.2	GDMs generated from the (a) AOD, (b) SIRP, and (c) SM datasets. For each plot, we display only those voxels that had an absolute z -score above 1.7. The red shows regions that are common to both the GDM extracted from the IVA-GL results as well as the GDM extracted from the ICA-EBM results. The yellow shows the regions that are in the GDM extracted from IVA-GL results but not in the GDM extracted from the ICA-EBM results. The cyan shows the regions that are in the GDM extracted from ICA-EBM results but not in the GDM extracted from the IVA-GL results.	84
-----	---	----

LIST OF TABLES

3.1	Number of significant components at two p -values, PCA-CCA estimated common components, d , and estimated order, C , for the three pairwise combinations of modalities. We assess statistical significance using a two-sample t -test run on the rows of the estimated subject covariations, $\hat{\mathbf{A}}^{[k]}, k = 1, 2$, calculated using tCCA.	40
5.1	Significance of the difference in classification rate using KSVM of ICA-EBM using a PCA order, C , of 15 compared to other values of C . We assess statistical significance through a two-sample t -test performed on the classification rates obtained for 200 independent resamplings of the fMRI feature datasets. If the difference corresponds to $p \geq 0.05$, the result is marked “not significant” (NS). Note that there is no correction for multiple comparisons and none of the observed differences would pass a Bonferroni correction (0.0056 for a p -value of 0.05).	65
5.2	Significance of the difference in classification rate using KSVM of IVA-GL compared to the data integration technique, combined ICAs, and the popular data fusion method, jICA. Statistical significance is assessed through a two-sample t -test performed on the classification rates obtained for 200 independent resamplings of the fMRI feature datasets. If the difference corresponds to $p \geq 0.05$, the result is marked “not significant” (NS). Note that there is no correction for multiple comparisons, but all significant differences would remain so after the conservative Bonferroni correction.	67

5.3	Significance of the difference in classification rate using KNN of IVA-GL compared with the data integration technique, combined ICAs, and the popular data fusion method, jICA. Statistical significance is assessed through a two-sample t -test performed on the classification rates obtained for 200 independent resamplings of the fMRI feature datasets. If the difference corresponds to $p \geq 0.05$, the result is marked “not significant” (NS). Note that there is no correction for multiple comparisons, but all significant differences would remain so after the conservative Bonferroni correction.	69
5.4	Significance of the difference in classification rate using FDA of IVA-GL compared with the data integration technique, combined ICAs, and the popular data fusion method, jICA. Statistical significance is assessed through a two-sample t -test performed on the classification rates obtained for 200 independent resamplings of the fMRI feature datasets. If the difference corresponds to $p \geq 0.05$, the result is marked “not significant” (NS). Note that there is no correction for multiple comparisons, but all significant differences would remain so after the conservative Bonferroni correction.	71
6.1	Significance in terms of p -values for each combination of dataset and analysis method measured using a 2-sample t -test run on the subject covariations of the GDM as well as Hotelling’s T -squared test. Note that the relative significance of each combination of dataset and analysis method is consistent across the two tests.	82

6.2 The first column from the left displays the t -statistic from the two-sample t -test run on the behavioral scores from the “measurement and treatment research to improve cognition in schizophrenia” consensus cognitive battery [1,2], where the first group is the scores from the controls and the second group is the scores from the patients. The second through seventh columns from the left are the correlation between the subject covariation from the GDM extracted from that combination of dataset and analysis method and the behavioral score. No correction for multiple comparisons was performed, however we highlight the entries that are not significant after a Bonferroni correction— $p < 2.02 \times 10^{-3}$ for a p -value of 0.05—in red. 84

LIST OF ABBREVIATIONS

AIC	Akaike's information criterion
AOD	auditory oddball
BIC	Bayesian information criterion
BSS	blind source separation
CCA	canonical correlation analysis
CCC	canonical correlation coefficients
EEG	electroencephalogram
ERP	event-related potential
FDA	Fisher discriminant analysis
fMRI	functional magnetic resonance imaging
GDM	global difference map
GENVAR	generalized variance
GLM	general linear model
HRF	hemodynamic response function
i.i.d.	independent and identically distributed
ICA	independent component analysis
ICA-EBM	entropy-bound minimization
ICA+tCCA	ICA then tCCA
ITC	information theoretic criteria
IVA	independent vector analysis
IVA-G	IVA with multivariate Gaussian prior

IVA-GGD	IVA with multivariate generalized Gaussian distribution prior
IVA-GL	IVA with multivariate Gaussian and Laplacian prior
IVA-L	IVA with multivariate Laplacian prior
JIVE	joint and individual variation explained
jICA	joint ICA
KNN	K -nearest neighbors
KSVM	kernel support vector machine
MAXVAR	maximum variance
MCCA	multiset CCA
MDL	minimum description length
MEG	magnetoencephalography
MINVAR	minimum variance
MST	minimum spanning tree
NPAIRS	nonparametric prediction, activation, influence, and reproducibility resampling
NS	not significant
p-ICA	parallel ICA
PCA	principal component analysis
PCA+CCA	PCA and CCA
PDF	probability density function
PET	positron emission tomography
SCM	source component matrix
SIRP	Sternberg item recognition paradigm
SM	sensory motor

sMRI	structural magnetic resonance imaging
SNP	single nucleotide polymorphism
SNR	signal-to-noise ratio
SPM	statistical parametric mapping toolbox
SSQCOR	maximum sum of squared correlation
SUMCOR	maximum sum of the correlation
SVD	singular value decomposition
tCCA	transposed CCA
tIVA	transposed IVA
tMCCA	transposed MCCA

Chapter 1

INTRODUCTION

1.1 Motivation

The collection of data from multiple sensors has become common in many fields, since these different sensors are expected to provide complementary views of complicated systems, such as the study of brain activity [3], which is the focus of this dissertation. We refer to data drawn from sensors of different types *modalities*. Thus, data drawn from multiple sensors can be categorized as either *multiset data* or *multimodal data*. Multiset data is drawn from sensors of the same type, such as: functional magnetic resonance imaging (fMRI) or electroencephalography (EEG) data from multiple tasks, at different times, or from different subjects or the red, green, and blue color channels in video sequences. Contrasted with multiset data, multimodal data is drawn sensors of different types, such as: fMRI data and EEG data from the same subjects and task, remote sensing, *i.e.*, light detection and ranging data as well as synthetic aperture data of the same region, and data from the diverse set of devices used in meteorological monitoring. This raises the issue of how to best utilize each of these datasets in order to maximize the use of available information for the given task. This goal is made more difficult by the fact that the datasets are generated from an intricate system, have fairly low signal-to-noise ratios (SNRs), and—in the case of multimodal data—have different physics units and are of different sizes [4]. In addition, due to the fact that the mechanisms underlying neural function are not fully understood, a

major challenge with such analyses is how to validate the results and assess/quantify the relative performance of different methods.

When applied to medical imaging, the methods used to analyze data from multiple sensors can be broadly classified as being one of two categories: *data integration* and *data fusion* [5, 6]. In data integration, the datasets are analyzed separately and the results are combined or the results of one technique are used to constrain another, see *e.g.*, [7–9]. This class of techniques is popular for the analysis of multimodal data, since, by analyzing each dataset separately, it alleviates the issue of jointly analyzing datasets of different dimensionality and physical units [4]. Data fusion methods, on the other hand, analyze the datasets together in a symmetric manner, enabling the datasets to fully interact and inform each other by exploiting the information shared across datasets, see *e.g.*, [6, 10–13]. The concept underlying data fusion is that an ensemble of related datasets can be more than the “sum of its parts,” meaning that it contains precious information, which can greatly improve the results if exploited and is lost if the relations across datasets are ignored. Though, potentially more powerful than data integration, a key issue for data fusion is the selection of a model that can faithfully represent the relationships between the datasets, thus explaining the similarities and differences between them, and yields a more interpretable combination of the datasets [4]. This is particularly an issue for multimodal data, since we would like to be able to define links between the heterogeneous datasets that are flexible enough to allow each dataset to remain in its most explanatory form.

Since, in general, little is known about the processes underlying brain function, it is important to reduce the potential for biases and therefore minimize the assumptions placed on the data. This motivates the use of blind source separation (BSS) techniques, of which independent component analysis (ICA) is perhaps the most popular [14, 15] for the fusion of medical imaging data drawn from multiple sensors, see *e.g.*, [12, 16–28]. BSS tech-

niques minimize the assumptions placed on the data through the use of a simple generative model, usually of the linear mixing type. However, each generative model assumes different underlying relationships between the datasets and using a model that is too restrictive or is unable to fully exploit the links between datasets will result in less meaningful or, in the most extreme case, incorrect results [4, 13, 29]. This incentivizes the development of BSS techniques with more general models that can describe a wider range of links between the datasets, such as independent vector analysis (IVA), a recent generalization of ICA to multiple datasets [30], for the fusion of both multiset and multimodal data. Despite the potential benefits that IVA offers, there are several issues that prevent its immediate adoption as the main method for fusion. These challenges include: determining the nature as well as the strength of the connection between the datasets, determining how these intrinsic properties interact with the assumptions of different analysis methods, and objectively evaluating the performance of different methods on real neuroimaging datasets.

1.2 Contributions

In this dissertation, we address the challenges that we discuss above, specifically: the determination of the strength of the connection between different datasets, the determination of how the intrinsic properties of the data interact with the assumptions of different analysis methods, and the determination of an unambiguous way to validate the results of different methods as well as assess/quantify their relative performance. In addressing these challenges, we make the following contributions:

- We introduce a new method to determine which multimodal neurological datasets are connected and to what degree. We investigate the performance of this technique using a series of simulations as well as multimodal neuroimaging data from three modalities and show how the proposed technique can be used in an exploratory anal-

ysis to assess which combinations of datasets have the greatest potential in a fusion analysis.

- We fuse multimodal neurological data from three modalities and demonstrate how this unique set of datasets enables the determination of the contribution of each dataset to the fusion result. We investigate how the strength of the connections across the datasets as well as practical concerns, such as the number of subjects, determine which BSS technique is most appropriate for the analysis.
- We propose a data-driven classification procedure to unambiguously quantify the performance of different analysis techniques as well as the amount of potentially exploitable information that a new dataset introduces to a decomposition beyond what already exists in the other datasets. We apply the proposed technique to multitask fMRI data to discover which datasets contribute most to the final result.
- We introduce a new method to summarize the amount of discriminative power for each dataset within a decomposition. We apply this method to multitask fMRI data and show how the proposed method facilitates comparisons across fusion techniques both in terms of the significance of the results as well as how they are related to certain behavioral scores.

1.3 Overview of Dissertation

The remainder of the dissertation is organized as follows. In Chapter 2, we provide an introduction to the data and the majority of the methods that we will use throughout this work. We begin by briefly introducing the relevant neuroimaging modalities that we analyze and describe how to use features in order to facilitate both multiset as well as multimodal fusion. We then discuss the implementation details of how we perform data

fusion, including: dimension-reduction, different BSS methods for fusion, and how we derive final results for comparison between multiple non-convex iterative methods.

In Chapter 3, we introduce the method to determine the strength of the connections between multimodal neurological datasets drawn from patients with schizophrenia and healthy controls. We show its superior performance over classical methods using simulations and apply it to real multimodal neuroimaging data from three modalities. We use the proposed method to determine which combinations of datasets share the strongest connection as well as which ones share the weakest.

In Chapter 4, we compare the performance of the techniques discussed in Chapter 2 on simulated neuroimaging data in order to test the robustness of the techniques to deviations from their assumptions. We also apply these methods to real multimodal neurological data drawn from patients with schizophrenia and healthy controls. We show how the strength of the connections across the datasets as well as practical concerns determine which BSS technique is most appropriate for the analysis. We also show how we can apply different methods on different combinations of the datasets to gain greater understanding about the connections between the modalities.

In Chapter 5, we introduce the classification procedure that is able to unambiguously quantify the performance of different BSS techniques on real multitask fMRI data drawn from patients with schizophrenia and healthy controls. We demonstrate how to use this method to determine the “value added” by each dataset to the analysis and show how each task contributes differently to the final fusion result.

In Chapter 6, we introduce a new visualization technique to highlight the changes from fusion and quantify the discriminatory power of each dataset within a decomposition. We apply this technique to real multitask fMRI data drawn from patients with schizophrenia as well as healthy controls and show how the technique facilitates comparisons across analysis

methods in terms of: the significance of the results, how the results are related to certain behavioral scores, and how the discriminatory brain regions change between methods.

We conclude the dissertation with Chapter 7 and present possible directions for future research.

Chapter 2

BACKGROUND

This chapter provides an introduction to the data and the majority of the methods that we will use throughout this dissertation. We begin by briefly introducing the relevant neuroimaging modalities that we analyze and describe how to use features in order to facilitate both multiset as well as multimodal fusion. We then discuss the implementation details of how we perform data fusion, including: dimension-reduction, different BSS methods for fusion, and how to derive results for comparison between multiple non-convex iterative methods.

2.1 Neuroimaging Modalities

The desire to understand how the brain encodes cognition and behavior has lead to the rise of many techniques to measure neural activity, with each exploiting a different aspect of neural function. These techniques include: fMRI, EEG, magnetoencephalography (MEG), near-infrared imaging, positron emission tomography (PET), and single photon emission computed tomography, among others. Additionally, since structure usually underlies function, a number of imaging techniques that solely study neural structure, such as: structural MRI (sMRI) and diffusion tensor imaging, have also been developed. Since each of these imaging techniques are different sensors that report on neural function, we refer to each of these different imaging techniques as different modalities.

Each of these modalities provides information about the brain at different spatial and temporal resolutions. For example, EEG and MEG have comparatively high temporal resolution, but low spatial resolution, meaning that these modalities can be used to study the instantaneous changes in neural activity, but they have difficulty determining the location in the brain where the activity occurred. Conversely, fMRI and PET have high spatial resolution, but have low temporal resolution, meaning that they can accurately locate the locations of neural activation, but they cannot gather this information quickly. Motivating a joint analysis of multimodal neuroimaging data is the desire to allow the complementary aspects of each modality, *i.e.*, their contrasting spatial and temporal resolutions, to augment each other and provide a more detailed view of functional changes within the brain. In this dissertation, we focus on the more commonly used and better understood functional neuroimaging techniques fMRI and EEG as well as the structural neuroimaging technique sMRI.

2.1.1 Feature Extraction

Due to the inherent dissimilarities between different modalities, as shown in Figure 2.1 for fMRI, sMRI, and EEG data, as well as the differences in timing for multitask fMRI data, direct fusion of multimodal and multisubject/multitask data is difficult. Therefore, rather than jointly analyzing the datasets directly, it is often beneficial to reduce each dataset to a feature for each subject within the study. This reduction using a lower-dimensional, yet still multivariate representation of the data facilitates exploration of the associations across these feature sets through the analysis of variations across individuals, see *e.g.*, [6, 31, 32]. Such an investigation of variations between subjects through fused features provides a natural way to discover associations across datasets and simplifies the identification of biomarkers of disease, since it constructs a common dimension for datasets of otherwise different dimensionality and in the case of multimodal data, as shown in Figure 2.1, different units as

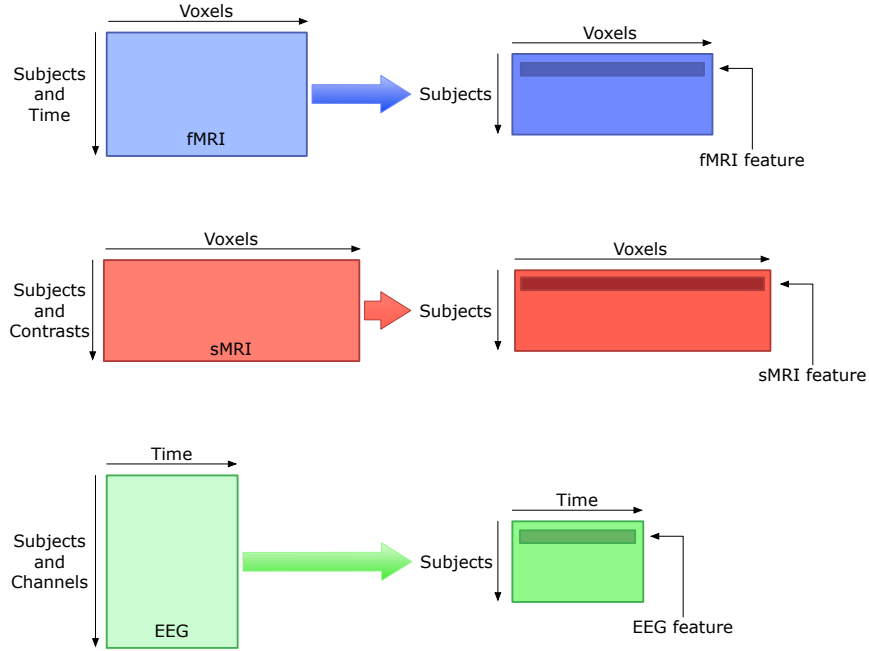


FIG. 2.1. Extraction of features for each subject from multimodal neuroimaging data. Note the differences in the dimensionality of the original data and how a common *subjects* dimension is established through the extraction of features from each subject.

well [12, 33]. In addition, as we describe in Section 2.3.3, this establishment of a common dimension and similarities between covariations across datasets as well as, especially in the case of multimodal fusion, dissimilarity across the datasets provide the motivation for the use of transposed IVA (tIVA). In the following sections, we describe the neuroimaging data used in this dissertation as well as the features that were extracted.

2.1.2 FMRI

The principle underlying the use of fMRI to study neural function is that brain activity, *i.e.*, neuronal activation, is associated with oxygen depletion in blood vessels near the activated neurons. This depletion is followed by an influx of oxygenated blood to the region surrounding the activated neurons. Since oxygenated blood has slightly different magnetic properties than deoxygenated blood, fMRI can provide an indirect and delayed measure

of brain function by recording these small changes in magnetic susceptibility. fMRI data consists of repeatedly imaging the 3-D volume of the brain in a series of slices, meaning that though fMRI data has fairly high spatial resolution, often at the millimeter scale, it suffers from poor temporal resolution, generally on the order of seconds. Since we expect that different tasks provide complementary views of neural function, in this dissertation, we also fuse fMRI data from multiple tasks. For each task and subject, we analyze the data from each voxel using a simple linear regression, using the general linear model (GLM), in the statistical parametric mapping toolbox (SPM), where the regressors are created by convolving the hemodynamic response function (HRF) in SPM with the desired predictors. Note that for these regressions, we assume that the noise is Gaussian and uncorrelated. We use the resulting regression coefficient maps as the feature for each fMRI task, described below.

Auditory Oddball (AOD) This auditory task involves subjects listening to three different types of auditory stimuli: standard (1 kHz tones occurring with probability 0.82), novel (computer generated, complex sounds occurring with probability 0.09), and target (1.2 kHz tones occurring with probability 0.09, to which a right thumb button press was required), in a pseudo-random order [34, 35]. Each run consists of 90 stimuli each with a 200 ms duration and a randomly changing interstimulus interval of 550–2,050 ms. The order of novel and target stimuli is changed between runs to ensure that the responses do not depend on the stimulus order [35]. In addition, the sequences of stimuli are designed such that they produce orthogonal blood-oxygen-level dependent responses for each of the three stimuli [36, 37]. For this task, the regressor is created by modeling the target and standard stimuli as delta functions convolved with the default SPM differences of gammas HRF in addition to their temporal derivatives [38]. We use subject-averaged contrast images between the target versus the standard tones as the feature for this task. We analyze data

from this task in Chapters 3, 4, 5, and 6.

Sternberg Item Recognition Paradigm (SIRP) In this visual task, the subjects have to remember a set of 1, 3, or 5 randomly chosen integers between 0 and 9. The task paradigm consists of: a 1.5 second *learn* condition, a blank screen for 0.5 seconds, a 6 second *encode* condition, where the sequence of digits is presented together, and a 38 second *probe* condition, where the subject is shown a series of integers and has to indicate, with a button press with the right thumb, whether it is a member of the memorized set or not [35]. Each probe digit is presented for up to 1.1 seconds in a pseudo-randomly jittered fashion within a 2.7 second interval [35]. A total of 84 probes, 42 targets and 42 foils are obtained per scan and the prompt-encode-probe conditions are run twice for each set size in a pseudo-random order [35]. For this task, the regressor is created by convolving the probe response block for the three digit set with the default SPM HRF [38]. This is done for both runs of the probe response and the average map is used as the feature for this task. We analyze data from this task in Chapters 5 and 6.

Sensory Motor (SM) In this auditory task, the subjects are presented with a sequence of auditory stimuli consisting of 16 different tones each lasting 200 ms and ranging in frequency from 236 Hz to 1,318 Hz, with a 500 ms inter-stimulus interval [35]. The first tone is set at the lowest pitch and each subsequent tone is higher than the previous one until the highest tone is reached and which point the order of the tones is reversed [35]. Each tonal change requires a button press with the right thumb. A total of 15 increase-and-decrease blocks are alternated with 15 fixation blocks, with each block lasting 16 seconds in duration [35]. For this task, the regressor is created by convolving the entire increase-and-decrease block with the default SPM HRF [38]. The responses for each subject are averaged over the separate runs and the average map is used as the feature for this task. We

analyze data from this task in Chapters 5 and 6.

2.1.3 EEG

Since neuronal activation is accomplished through the movement of charged ions between neurons and the surrounding environment, a direct measure of brain activity can be achieved by recording electrical activity through electrodes placed on the scalp. This enables a more direct measure of neural activity compared with fMRI and the signal can be recorded almost instantaneously, most commonly at the millisecond scale. However, EEG has poor spatial resolution, often at the centimeter scale, and determining the exact location of the neural activation is a challenging problem. This problem is made even more difficult due to the fact that the electrodes are more sensitive to neural activation occurring closer to the scalp than that occurring deep within the brain. The EEG data that we analyze in this dissertation was collected during the performance of the AOD task described previously. By averaging timelocked windows around the target tone over the repeated instances of the stimuli, event-related potentials (ERPs) can be extracted from each electrode and generally offer higher SNR than the raw EEG. In this dissertation, we use the ERP corresponding to the midline central position (Cz) channel as the feature for fusion, due to the fact that it appears to be the best single channel to detect both anterior and posterior sources. Also, when displaying the ERP components, we always compare them with the subject-averaged raw ERP. We analyze data from this modality in Chapters 3 and 4.

2.1.4 SMRI

SMRI describes the neural structure of the brain through identification of the locations of the gray matter, white matter, and cerebrospinal fluid. This is done by exploiting the different properties of these three media by varying different parameters of the MRI machine, such as: repetition time, echo time, or flip angle. Since each scan uses a variety of pa-

parameter values, the spatial resolution is very high, generally, at the sub-millimeter scale, however, there is no temporal information. In this dissertation, we extract probabilistically segmented gray matter images as the sMRI feature for fusion. We analyze data from this modality in Chapters 3 and 4.

2.2 Dimension-Reduction and Order Selection

Since medical imaging data is usually quite noisy and of high dimensionality, *i.e.*, is generated by a smaller number of factors than subjects and has a relatively large number of data samples, dimension-reduction, most commonly using principal component analysis (PCA), is a crucial preprocessing step for avoiding the problem of over-fitting in subsequent analyses. However, the success of this step is intimately tied the concept of order-selection, namely, determination of how much of the data to retain. Note that different fusion models exploit different dimensions of the data in order to achieve a decomposition, namely, joint ICA (jICA), which we describe in Section 2.3.1 and IVA, which we describe in Section 2.3.2, exploit the voxels/time dimension whereas tIVA, which we describe in Section 2.3.3, exploits the subjects dimension.

The basic assumption underlying classical order-selection methods is that the signals of interest reside in a lower dimensional “signal subspace” and that the entire dataset is corrupted by isotropic white Gaussian noise. Because of this, the dimension of the signal subspace, *i.e.*, the order, can be found, by assuming the data is multivariate Gaussian distributed or the majority of the energy of the signals resides in the second-order moments, by looking at the eigenvalues of the sample covariance matrix. Then, the data can be projected into the signal subspace by projecting the data onto the subspace spanned by the eigenvectors corresponding to the highest eigenvalues. Though simple in theory, this task is often more difficult in practice, since there is often no clear distinction between

the eigenvalues associated with the signal subspace and the eigenvalues associated with the noise subspace. For a single dataset, the most popular order-selection methods, such as: Akaike's information criterion (AIC) [39], minimum description length (MDL) [40] or, equivalently, Bayesian information criterion (BIC) [41], and extensions of those methods, see *e.g.*, [42], define the order based on information theoretic criteria (ITC) [43], *i.e.*, by using a function of the estimated eigenvalues of the covariance matrix of the data and the number of model parameters as shown in (2.3) and described next.

Given a feature dataset from M subjects, $\mathbf{X} = [\mathbf{x}(1), \dots, \mathbf{x}(V)] \in \mathbb{R}^{M \times V}$, where the m th row of \mathbf{X} is formed either by flattening the feature of V voxels from the m th subject for fMRI and sMRI data or simply taking the V feature time points from the m th subject for EEG data, the GLM for \mathbf{X} can be written as

$$\mathbf{X} = \mathbf{A}\mathbf{S}, \quad (2.1)$$

where the $0 \leq C < M$ latent sources or neural spatial/temporal activation patterns, $\mathbf{S} = [\mathbf{s}_1, \dots, \mathbf{s}_C]^T \in \mathbb{R}^{C \times V}$, are linearly mixed by the mixing matrix, $\mathbf{A} \in \mathbb{R}^{M \times C}$. The general form of the ITC for \mathbf{X} is given by

$$\widehat{C} = \underset{k}{\operatorname{argmin}} \{ \alpha \mathcal{L}(\mathbf{X}|C) + r(\theta_k) \eta(V) \}, \quad (2.2)$$

where $\mathcal{L}(\mathbf{X}|C) = -\log P(\mathbf{x}(1), \dots, \mathbf{x}(V)|\theta_k)/V$ is the normalized negative log-likelihood of \mathbf{X} , given that model order is k , and $r(\theta_k)$ indicates the number of free parameters, given that the model order is k . Different methods use different values for α and $\eta(V)$. For example, AIC uses $\alpha = 2$ and $\eta(V) = 2/V$, whereas MDL and BIC use $\alpha = 1$ and $\eta(V) = (\log V)/2V$. Note that for the model given in (2.1) with independent and identically distributed (i.i.d.) samples, order detection using MDL is consistent as $V \rightarrow \infty$ and the AIC is inconsistent [43]. In the subsequent discussion, we focus on MDL, but using the appropriate expressions

for α and $\eta(V)$, the form can be extended to other criteria.

For the model given in (2.1) with i.i.d. samples, the MDL order-selection criterion [42–44] can be written as

$$\widehat{C} = \underset{k}{\operatorname{argmin}} \left\{ \sum_{i=1}^k \log \lambda_i^2 + (M - k) \log \sigma^2 + \frac{k(2M - k + 1) \log V}{4V} \right\}, \quad (2.3)$$

where λ_i is the i th largest eigenvalue of the sample covariance matrix of \mathbf{X} , $\hat{\mathbf{R}}_{\mathbf{X}\mathbf{X}} = \mathbf{X}\mathbf{X}^T/V$, which indicates the energy of the i th principal component, and $\sigma^2 = (M - k)^{-1} \sum_{i=k+1}^M \lambda_i^2$, is the total energy of the noise subspace. The popularity of ITC-based order-selection methods stems from the fact that the order can be consistently estimated as the sample size increases to infinity, without requiring the user to specify subjective thresholds, as is the case for approaches based upon hypothesis-testing or other *ad hoc* techniques, see *e.g.*, [45, 46]. We further discuss order-selection and the development of methods that work in the high noise of neuroimaging data and, in the case of multimodal data, the sample-poor regime in Chapter 3.

2.3 BSS Techniques for Data Fusion

Neuroimaging data analysis techniques can be broadly categorized as being model-based, data-driven, or a hybrid between the two, see *e.g.*, [47–49]. Model-based techniques seek to model the experimental design, generally using a model of some neurological process, such as the HRF for fMRI or source localization for EEG, usually through a GLM-based approach, see *e.g.*, [50–53]. However, the assumptions do not always hold true in practice, especially across large populations affected by different neurological disorders [54], weakening the conclusions that can be drawn from these analyses. The desire to generate meaningful summary factors from the data, while reducing potential biases, through the minimization of the assumptions placed on the data, has lead to the popularity of data-

driven fusion methods, generally based on BSS, for the fusion of neuroimaging data. These fusion methods include: group ICA [16], jICA [17, 55], partial least squares [18], parallel factors/canonical polyadic decomposition [20], tensor ICA [21], canonical correlation analysis (CCA) [22], multiset CCA (MCCA) [12], IVA [23, 24], parallel ICA (p-ICA) [25], coefficient constrained ICA [26], CCA/MCCA+jICA [27, 28], linked ICA [56], constrained matrix-tensor factorization [57], block term decomposition [58], among others. For a more in depth discussion of different fusion methods, their uses, and assumptions, see [4, 33]. In the subsequent sections, we describe the three classes of techniques that we use in this dissertation, namely jICA, IVA, and tIVA, with an emphasis on their connections.

2.3.1 JICA

Returning to the generative model in (2.1), it is clear that \mathbf{A} and \mathbf{S} are not unique, since right-multiplying \mathbf{A} by an invertible $C \times C$ matrix \mathbf{Q} and left-multiplying \mathbf{S} by \mathbf{Q}^{-1} will result in another, equally valid, solution. Since it is usually assumed that the number of subjects is larger than the number of latent sources, the first step in factoring \mathbf{X} is performing dimension-reduction, where the dimension of the signal subspace has been determined previously. This reduction is generally performed using PCA, thus reducing \mathbf{X} to $\underline{\mathbf{X}}$ by

$$\underline{\mathbf{X}} = \mathbf{F}\mathbf{X}, \quad (2.4)$$

where \mathbf{F} is the dimension-reduction matrix and is equal to the eigenvectors corresponding to the \widehat{C} highest eigenvalues of the sample covariance matrix of \mathbf{X} and \widehat{C} is estimated using (2.2). Following this reduction and by assuming independence on the part of the latent sources as well as that the dimension-reduced mixing matrix, $\underline{\mathbf{A}} = \mathbf{F}\mathbf{A}$, is full rank, ICA can estimate an essentially unique demixing matrix, \mathbf{W} , *i.e.*, subject to only scaling and permutation ambiguities, such that the estimated components, $\hat{\mathbf{S}} = [\hat{\mathbf{s}}_1, \dots, \hat{\mathbf{s}}_{\widehat{C}}]^T$, are given

by

$$\hat{\mathbf{S}} = \mathbf{W}\underline{\mathbf{X}}, \quad (2.5)$$

and the estimated mixing matrix, $\hat{\mathbf{A}}$, is given by

$$\hat{\mathbf{A}} = \mathbf{F}^T \hat{\mathbf{A}} = \mathbf{F}^T \mathbf{W}^{-1}. \quad (2.6)$$

This maximization of independence can be formulated using a maximum likelihood framework, assuming that the samples, *i.e.*, columns, of $\underline{\mathbf{X}}$, $\underline{\mathbf{x}}(v)$, $v = 1, \dots, V$, are i.i.d., as

$$\begin{aligned} \mathcal{L}_{\text{ICA}}(\mathbf{W}) &= \log(p_{\underline{\mathbf{X}}}(\underline{\mathbf{X}})) \\ &= \log\left(\prod_{v=1}^V p_{\underline{\mathbf{X}}}(\underline{\mathbf{x}}(v))\right) \\ &= \sum_{v=1}^V \log(p_{\mathbf{S}}(\mathbf{W}\underline{\mathbf{x}}(v)) |\det(\mathbf{W})|) \\ &= \sum_{v=1}^V \log(p_{\mathbf{S}}(\mathbf{W}\underline{\mathbf{x}}(v))) + V \log(|\det(\mathbf{W})|) \end{aligned} \quad (2.7)$$

$$= \sum_{v=1}^V \left[\sum_{c=1}^{\hat{C}} \log p_{\hat{s}_c}(\mathbf{w}_c^T \underline{\mathbf{x}}(v)) \right] + V \log |\det(\mathbf{W})|, \quad (2.8)$$

where \mathbf{w}_c^T is the c th row of \mathbf{W} and $p_{\hat{s}_c}(\mathbf{w}_c^T \underline{\mathbf{x}}(v))$ is the probability density function (PDF) from which we have the samples \hat{s}_c , *i.e.*, the c th estimated component. Note that the transition from (2.7) to (2.8) is done through the assumption of independence on the part of the latent sources. Using the mean ergodic theorem and as $V \rightarrow \infty$, maximization of (2.8) can be shown to be equivalent to the minimization of the mutual information between the random variables $p_{\hat{s}_c}(\mathbf{w}_c^T \underline{\mathbf{x}}(v))$ [59].

Since the rows of the dataset are composed of features that we extract from each subject, the columns of the estimated mixing matrix, $\hat{\mathbf{A}}$, provide the loadings of each component across subjects. Thus, the m th column of the estimated mixing matrix, $\hat{\mathbf{a}}_m$, represents the relative weights of the m th estimated component, \hat{s}_m , for each corresponding subject,

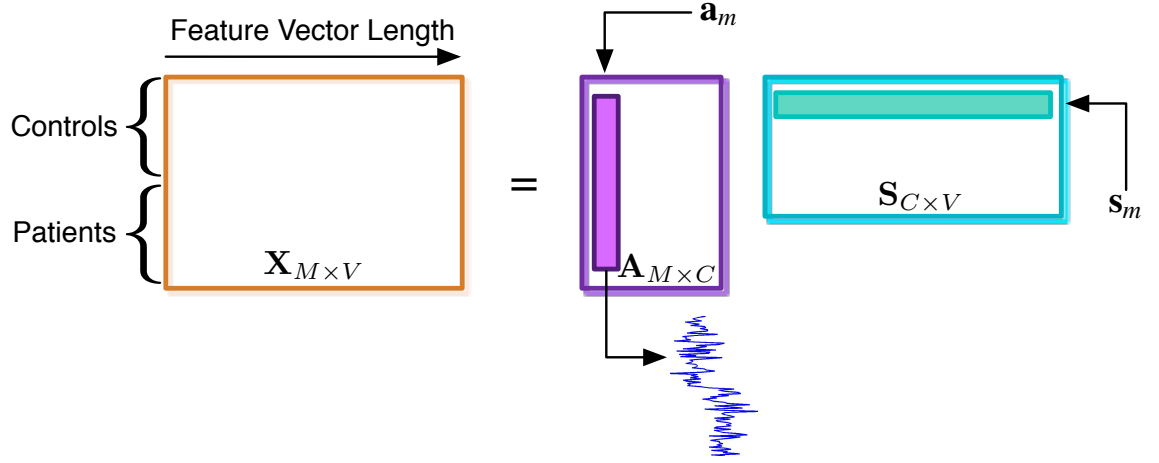


FIG. 2.2. Generative model for ICA. Note how the m th column of the mixing matrix, \mathbf{a}_m , represents the relative weights of the m th source, \mathbf{s}_m , for each corresponding subject.

as shown in Figure 2.2. Therefore, to look for differences in the expression of components between two groups, we can perform a two-sample t -test on the subject covariations, where one group is represented by the subject covariations from the patients with schizophrenia and the other by the subject covariations from the healthy controls [6]. We often refer to those components for which the two-sample t -test on the corresponding subject covariations was statistically significant as potential biomarkers of disease. Note that the sign of the t -statistic determines whether the component has higher association in the patients with schizophrenia or the healthy controls. In this dissertation, we display only the components that have corresponding t -statistics that are positive, *i.e.*, that have higher activation in the controls versus the patients. If a component is a biomarker, but has a negative t -statistic, we multiply both the component and the corresponding column of the mixing matrix by -1 in order to make the t -statistic positive. It is important to note that this does not change the fundamental nature of the component, due to the sign ambiguity inherent to ICA.

There are many different ICA algorithms, each with different assumptions about the PDF of the latent sources. The underlying PDF of the latent sources is usually unknown

in most applications and ICA algorithms that utilize a fixed or simple model for the latent source PDF will result in poor separation performance when the data deviates significantly from the assumed model. In this dissertation, we use the entropy-bound minimization (EBM) algorithm, due to the fact that it shows superior performance in both simulated and real neurological data when compared with the popular Infomax algorithm [60], see *e.g.*, [61–63]. This increase in performance is due to the fact ICA-EBM does not assume any distribution for the latent sources and instead attempts to upper bound their entropy through the use of several measuring functions [61]. These measuring functions describe a wide variety of distributions, including those that are: unimodal, bimodal, symmetric, and skewed [61], thus generally leading to accurate estimation of all sources within the mixture. Infomax, on the other hand, through the use of a fixed sigmoidal nonlinearity, is equivalent to assuming that the latent sources are distributed according to the hyperbolic secant distribution using a maximum likelihood framework [64]. This super-Gaussian source density model is a good match to focal, *i.e.*, sparse, fMRI sources, see *e.g.*, [65], however it can be less successful with less focal fMRI sources [66] as well as the “joint” sources found using jICA [67], which we describe next.

We extend the model in (2.1) in a straightforward manner to K datasets as

$$\mathbf{X}^{[k]} = \mathbf{A}^{[k]} \mathbf{X}^{[k]}, \quad 1 \leq k \leq K. \quad (2.9)$$

Due to the inherent scaling and permutation ambiguities of ICA, running ICA individually on each dataset and attempting to align the results is impractical and suboptimal, particularly as K increases. For this reason and to exploit interactions across different modalities/tasks, it is possible to assume that each dataset is mixed with the same mixing matrix, \mathbf{A} , which enables the solution of the problem posed in (2.9) through the performance of a single ICA in which the sources from the disparate datasets form underlying

“joint sources” [19, 31]. We can then jointly estimate the sources, and hence, determine the individual components, through the performance of a single ICA on the horizontally concatenated $\mathbf{X}^{[k]}$ defined as

$$\begin{aligned}\bar{\mathbf{X}} = [\mathbf{X}^{[1]}, \mathbf{X}^{[2]}, \dots, \mathbf{X}^{[K]}] &= \mathbf{A}[\mathbf{S}^{[1]}, \mathbf{S}^{[2]}, \dots, \mathbf{S}^{[K]}] \\ &= \mathbf{A}\bar{\mathbf{S}}.\end{aligned}\tag{2.10}$$

This method, referred to as jICA [55], is one of the most popular data fusion methods for multisubject/multitask as well as multimodal neuroimaging data, see *e.g.*, [11, 17, 27, 31, 68–73]. However, jICA is reliant on the assumption that each dataset has the same mixing matrix, thus it may perform poorly when this is not the case [13]. In addition, jICA inherently values each dataset by its dimensionality, meaning that it would weigh datasets with a lower dimensionality, such as EEG, less than those with a higher dimensionality, such as fMRI and sMRI.

2.3.2 CCA, MCCA, and IVA

An alternative to jICA for two datasets is CCA [74] or its extension to multiple datasets, MCCA [75], which both maximize the correlation of sources across datasets. Neither CCA nor MCCA restrict their solution space by assuming that each mixing matrix is identical, but rather estimate a separate set of mixing matrices, $\hat{\mathbf{A}}^{[k]}$, for each dataset simultaneously. Because of their less restrictive forms, CCA, MCCA, and related methods have successfully been applied to multisubject analyses of medical imaging data, see *e.g.*, [76–80] and to the fusion of multitask fMRI data [27].

CCA seeks to transform two datasets in order to maximize the correlation of the transformed datasets. This can be expressed in the following maximization problem, where in the first stage, we obtain the first rows of the transformation matrices $\mathbf{P}^{[1]}$ and $\mathbf{P}^{[2]}$, $\mathbf{p}_1^{[1]}$ and

$\mathbf{p}_1^{[2]}$, respectively,

$$\max_{\mathbf{p}_1^{[1]}, \mathbf{p}_1^{[2]}} \text{corr}(\mathbf{p}_1^{[1]} \underline{\mathbf{X}}^{[1]}, \mathbf{p}_1^{[2]} \underline{\mathbf{X}}^{[2]}) = \frac{\mathbf{p}_1^{[1]} \hat{\mathbf{R}}_{\underline{\mathbf{X}}^{[1]} \underline{\mathbf{X}}^{[2]}} (\mathbf{p}_1^{[2]})^T}{\sqrt{\mathbf{p}_1^{[1]} \hat{\mathbf{R}}_{\underline{\mathbf{X}}^{[1]} \underline{\mathbf{X}}^{[1]}} (\mathbf{p}_1^{[1]})^T \mathbf{p}_1^{[2]} \hat{\mathbf{R}}_{\underline{\mathbf{X}}^{[2]} \underline{\mathbf{X}}^{[2]}} (\mathbf{p}_1^{[2]})^T}}, \quad (2.11)$$

where $\hat{\mathbf{R}}_{\underline{\mathbf{X}}^{[i]} \underline{\mathbf{X}}^{[j]}} = \underline{\mathbf{X}}^{[i]} (\underline{\mathbf{X}}^{[j]})^T / V$ denotes a sample (cross-) covariance matrix and the operation $\text{corr}(\cdot, \cdot)$ refers to the sample correlation between the projected datasets. We can estimate subsequent rows of $\mathbf{P}^{[1]}$ and $\mathbf{P}^{[2]}$ analogously to (2.11), with the additional constraint that

$$\mathbf{p}_i^{[k]} \underline{\mathbf{X}}^{[k]} (\mathbf{p}_j^{[k]} \underline{\mathbf{X}}^{[k]})^T = 0, \quad \forall j \neq i, \quad k = 1, 2.$$

The estimated sources are then given by,

$$\hat{\mathbf{S}}^{[k]} = \mathbf{P}^{[k]} \underline{\mathbf{X}}^{[k]}, \quad k = 1, 2, \quad (2.12)$$

and estimated mixing matrices by

$$\hat{\mathbf{A}}^{[k]} = (\mathbf{F}^{[k]})^T (\mathbf{P}^{[k]})^T, \quad k = 1, 2. \quad (2.13)$$

MCCA extends CCA to multiple datasets and maximizes correlation across the datasets by trying to make the covariance matrices, $\mathbf{R}_{\hat{\mathbf{S}}_c \hat{\mathbf{S}}_c}$, of each source component matrix (SCM), defined as $\hat{\mathbf{S}}_c = [\hat{\mathbf{s}}_c^{[1]} \dots \hat{\mathbf{s}}_c^{[K]}]^T \in \mathbb{R}^{K \times V}$, *i.e.*, by concatenating the c th component from each of the datasets, as ill-conditioned as possible. This somewhat *ad hoc* technique to maximize the correlation across multiple datasets can be performed using one of five cost functions that all reduce to CCA when $K = 2$ [75]. These functions are:

- Maximum variance (MAXVAR): Maximize the largest eigenvalue of $\mathbf{R}_{\hat{\mathbf{S}}_c \hat{\mathbf{S}}_c}$
- Minimum variance (MINVAR): Minimize the smallest eigenvalue of $\mathbf{R}_{\hat{\mathbf{S}}_c \hat{\mathbf{S}}_c}$
- Minimum generalized variance (GENVAR): Minimize the determinant of $\mathbf{R}_{\hat{\mathbf{S}}_c \hat{\mathbf{S}}_c}$

- Maximum sum of the correlation (SUMCOR): Maximize the sum of the entries of $\mathbf{R}_{\hat{\mathbf{s}}_c \hat{\mathbf{s}}_c}$
- Maximum sum of squared correlation (SSQCOR): Maximize the sum of the squared entries of $\mathbf{R}_{\hat{\mathbf{s}}_c \hat{\mathbf{s}}_c}$.

These five cost functions are closely related. Defining λ_{\max} and λ_{\min} to be the largest and smallest eigenvalues of $\mathbf{R}_{\hat{\mathbf{s}}_c \hat{\mathbf{s}}_c}$, respectively, we see that the maximum of the SSQCOR cost, *i.e.*, the maximum of $\sum_i \lambda_i^2$, is dominated by λ_{\max} [75]; similarly, we see that the minimum of the GENVAR cost, *i.e.*, the minimum of $\prod_i \lambda_i$, is dominated by λ_{\min} [75]. Also, note that SUMCOR and SSQCOR are equivalent for correlation maximization [81].

However, despite their widespread use, CCA and MCCA have two important limitations that can hurt their applicability to real data: first, both CCA and MCCA only exploit second-order statistics and second, both CCA and MCCA limit the solution space by assuming that the demixing matrices are orthogonal. A way to overcome the shortcomings of CCA and MCCA is through the use of IVA, a recent multiset extension of ICA, which exploits similarities across datasets to achieve a successful decomposition. Unlike CCA or MCCA, IVA does not limit the demixing matrices to be orthogonal and can take advantage of second as well as higher-order statistics [59]. This flexibility has lead to the use of IVA in multisubject fMRI analyses, see *e.g.*, [23, 82, 83], however, to our knowledge, it has not yet been applied to the problem of fusing multitask feature data.

Given the definition of an SCM from MCCA, we can write the maximum likelihood objective function for IVA as

$$\mathcal{L}_{\text{IVA}}(\mathcal{W}) = V \left(\frac{1}{V} \sum_{v=1}^V \left[\sum_{c=1}^{\hat{C}} \log p_{\hat{\mathbf{s}}_c}(\hat{\mathbf{s}}_c(v)) \right] + \sum_{k=1}^K \log |\det(\mathbf{W}^{[k]})| \right), \quad (2.14)$$

where $\mathcal{W} = \{\mathbf{W}_1, \dots, \mathbf{W}_K\}$ and $p_{\hat{\mathbf{s}}_c}$ denotes the K -dimensional PDF from which the samples corresponding to the c th SCM are drawn. Similar to the case for ICA, it has been

shown that the maximum likelihood objective is asymptotically equivalent the mutual information objective function, with the difference being that we are now minimizing the mutual information between SCMs and not components [59]. Rewriting 2.14, we have

$$\mathcal{L}_{\text{IVA}}(\mathcal{W}) = V \left(\frac{1}{V} \sum_{v=1}^V \left[\sum_{c=1}^{\hat{C}} \left(\sum_{k=1}^K \log p_{\hat{s}_c^{[k]}}(\hat{s}_c^{[k]}(v)) - \log p_{\hat{s}_c}(\hat{s}_c(v)) \right) \right] + \sum_{k=1}^K \log |\det(\mathbf{W}^{[k]})| \right), \quad (2.15)$$

The new term, $\log p_{\hat{s}_c}(\hat{s}_c(v))$, in (2.15) represents the information between sources from different datasets within the SCM and is maximized by maximizing the negative log-likelihood function. Thus, we see how IVA exploits the shared information across datasets, when it exists, since if the term is not present, the objective function (2.15) is equivalent to performing ICA on each dataset separately [59].

Like ICA, there exists a variety of IVA algorithms that make different assumptions about the generative model of the latent sources, *i.e.*, the SCMs. The models include: the multivariate Laplacian, resulting the IVA-L algorithm [30], the multivariate Gaussian, resulting in the IVA-G algorithm [84], and the multivariate generalized Gaussian, resulting in the IVA-GGD algorithm [85]. An interesting note is that the IVA-G algorithm reduces to MCCA using the GENVAR cost if we assume that the $\mathbf{W}^{[k]}$ are orthogonal [13], thus IVA can be seen as a generalization of both CCA and MCCA.

2.3.3 TIVA

The previously described CCA/IVA models exploit the similarities of the spatial brain regions or temporal patterns in multi-subject/multitask fMRI or EEG data, respectively. However, we cannot use those techniques for multimodal fusion, due to the difficulty of evaluating the correlation term in (2.11) or the information term in (2.15) for dissimilar modalities. In addition, even if it were mathematically tractable to solve these issues, the results would not be meaningful, since increasing the similarity among datasets of fun-

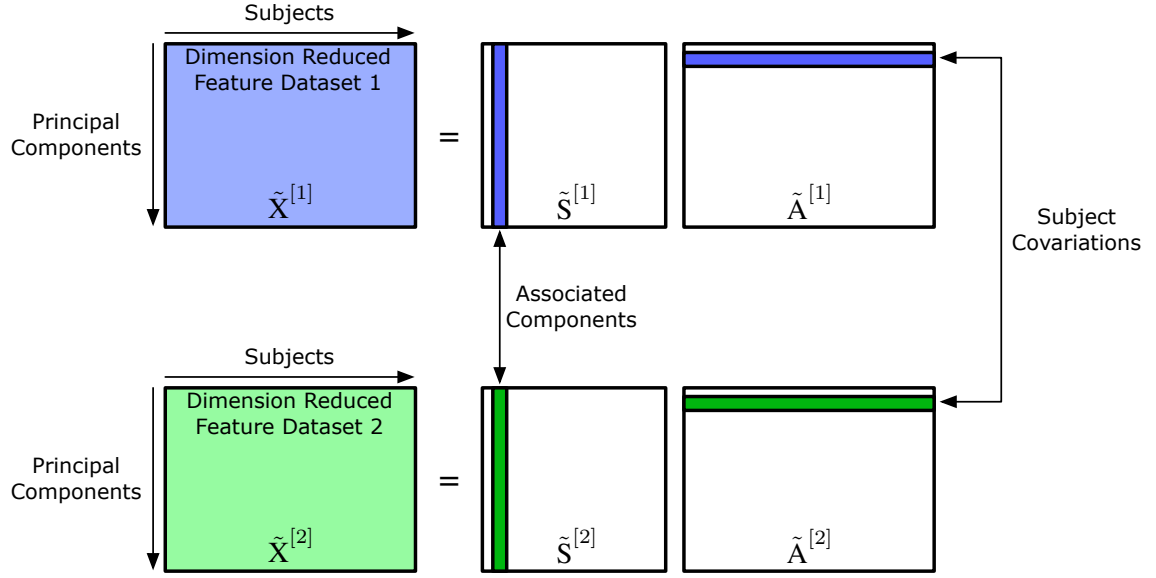


FIG. 2.3. Generative model for multimodal fusion using tCCA. Note that this model is obtained after spatial dimension-reduction through PCA, *i.e.*, $\tilde{\mathbf{X}}^{[1]}$ and $\tilde{\mathbf{X}}^{[2]}$ are both $C \times M$ matrices.

damentally different natures and units has no physical interpretation. Instead, the subject covariations provide the dimension of coherence in multimodal fusion using CCA/IVA, since similarity between covariations across modalities is expected in practice. As will be seen in the subsequent discussion, we refer to these models, which are quite useful for the fusion of multimodal neurological data, see *e.g.*, [12, 28, 73], as the transposed CCA (tCCA) and transposed IVA (tIVA) models, in order to differentiate them from their multi-subject/multitask versions. Figure 2.3 depicts the model for multimodal fusion using tCCA following spatial dimension-reduction through PCA; the implementation of tCCA for fusion is presented next. Note that we only present the discussion for tCCA, since tIVA is a straightforward extension of the methods used for tCCA.

Consider the transpose of two sets of multimodal feature data $(\mathbf{X}^{[1]})^T$ and $(\mathbf{X}^{[2]})^T$ gathered from the same M subjects, whose dimensions are $V_1 \times M$ and $V_2 \times M$. In this dissertation, these datasets could be any pairwise combination of the fMRI, sMRI, and EEG

data.

Prior to performing multimodal data fusion using tCCA, spatial dimension-reduction is first performed using PCA, in order to reduce $(\mathbf{X}^{[1]})^T$ and $(\mathbf{X}^{[2]})^T$ to $\tilde{\mathbf{X}}^{[1]}$ and $\tilde{\mathbf{X}}^{[2]}$, respectively, such that

$$\tilde{\mathbf{X}}^{[k]} = \mathbf{G}^{[k]} (\mathbf{X}^{[k]})^T, \quad k = 1, 2, \quad (2.16)$$

where $\mathbf{G}^{[k]}$ are data-reduction matrices and are equal to the eigenvectors corresponding to the highest C eigenvalues of the sample covariance matrix of $(\mathbf{X}^{[k]})^T$. Note that both $\tilde{\mathbf{X}}^{[1]}$ and $\tilde{\mathbf{X}}^{[2]}$ have the same dimensionality, *i.e.*, $C \times M$. We can decompose these datasets into the combination of subject covariations, $\tilde{\mathbf{A}}^{[1]}$ and $\tilde{\mathbf{A}}^{[2]}$, and associated components, $\tilde{\mathbf{S}}^{[1]}$ and $\tilde{\mathbf{S}}^{[2]}$, in the following manner,

$$\tilde{\mathbf{X}}^{[k]} = \tilde{\mathbf{S}}^{[k]} \tilde{\mathbf{A}}^{[k]}, \quad k = 1, 2. \quad (2.17)$$

By using this decomposition, we can evaluate the relationship across modalities based on the correlation of the subject covariations between the two modalities. We assume that the subject covariations are uncorrelated within each modality and are correlated with at most one subject covariation from the other modality. This enables the exploration of the associations across the modalities through relations between the subject correlations by performing CCA on $\tilde{\mathbf{X}}^{[1]}$ and $\tilde{\mathbf{X}}^{[2]}$, which produces the estimated subject covariations, $\hat{\mathbf{A}}^{[1]}$ and $\hat{\mathbf{A}}^{[2]}$, according to (2.12) [22].

Then, we find the estimated dimension-reduced sources, $\hat{\tilde{\mathbf{S}}}^{[1]}$ and $\hat{\tilde{\mathbf{S}}}^{[2]}$, using (2.17). Finally, we can find the full estimated sources, $\hat{\mathbf{S}}^{[1]}$ and $\hat{\mathbf{S}}^{[2]}$, through back-reconstruction of $\hat{\tilde{\mathbf{S}}}^{[1]}$ and $\hat{\tilde{\mathbf{S}}}^{[2]}$ using

$$\hat{\mathbf{S}}^{[k]} = (\mathbf{G}^{[k]})^T \hat{\tilde{\mathbf{S}}}^{[k]}, \quad k = 1, 2. \quad (2.18)$$

Note that, in principle, we can use the same algorithms used for MCCA and IVA for tMCCA and tIVA. However, since usually $M \ll V$ in practice, some care must be taken

when selecting an algorithm to use, especially for tIVA, due to the issues associated with fitting a complex model with a severely limited number of samples.

2.4 Minimum Spanning Tree (MST)

Since, aside from CCA and MCCA using the MAXVAR cost, all of the algorithms that we use in this dissertation are iterative in nature and the existence of local optima, there is a natural need to determine a result that best represents the average decomposition of a given set of data for use when comparing the fusion results from different algorithms. For this reason, we run each iterative algorithm ten times and identify what we call the *central run* using a MST-based method [66] and use that for further analyses and comparisons with other techniques. We use this MST-based method due to the fact that it shows superior performance to the popular alternative method of ICASSO [86]. The MST-based method works by first generating a matrix for each pair of runs whose elements are given by 1 minus the absolute value of the correlation coefficient between the estimated components from those runs. We find the minimum cost of aligning the sources between runs using the Hungarian algorithm [87], and construct a graph with nodes corresponding to runs and edges corresponding to the minimum cost of aligning the runs. Then, we generate an MST, with corresponding central node, for this graph. Since the method was originally proposed for ICA, we can directly apply it to the spatially concatenated “joint” components for jICA. For the IVA and tIVA models, we first align the components across datasets based on the sign of the correlation between each pair of corresponding components, such that whenever the correlation is negative for a given dataset, then both the estimated component and the corresponding subject covariation are multiplied by -1 .

2.5 Summary

Different neuroimaging modalities and different tasks within a single modality highlight unique aspects of neural processing, thus motivating their joint use in a single analysis. However, there are a number of issues associated with performing such an analysis including: varying dataset dimensionality and, in the case of multimodal data, different physical units. We can resolve these issues through the extraction of features from each dataset and exploring the associations across the feature datasets. In order to avoid possibly biasing the results, we seek to minimize the assumptions placed on the data and data-driven methods provide a promising route to analyze such data. However, there are number of data-driven methods that can be used, each based on different assumptions about the form of the latent sources. In addition, many of these techniques are iterative in nature, thus necessitating the selection of a run that represents a consistent estimate for use in further analyses.

Chapter 3

ORDER-SELECTION FOR MULTIMODAL DATA

Due to their data-driven nature, BSS methods, such as tCCA, are very useful for the fusion of multimodal neurological data. However, being able to determine the degree of similarity between datasets and appropriate order-selection are crucial to the success of such techniques. The standard methods for calculating the order of multimodal data focus only on sources with the greatest individual energy and ignore relations across datasets. Additionally, these techniques as well as the most widely-used methods for determining the degree of similarity between datasets assume sufficient sample support and are not effective in the sample-poor regime. In this chapter, we discuss joint estimation of the degree of similarity between datasets and their order when few samples are present using the method PCA and CCA (PCA-CCA). By considering these two problems simultaneously, we are able to minimize the assumptions placed on the data and achieve superior performance in the sample-poor regime compared to traditional techniques. We apply PCA-CCA to the pairwise combinations of multimodal data: fMRI, sMRI, and EEG data drawn from patients with schizophrenia and healthy controls while performing an auditory oddball task. The PCA-CCA results indicate that the fMRI and sMRI datasets are the most similar, whereas the sMRI and EEG datasets share the least similarity. We also demonstrate that the degree of similarity that we obtain using PCA-CCA is highly predictive of the degree of significance found for components generated using tCCA and therefore relevant to multimodal

fusion methods that exploit the tIVA model that we describe in Chapter 2.3.3.

3.1 Introduction

Despite being an effective multivariate and data-driven fusion method, see *e.g.*, [13, 22, 27, 73], CCA requires sufficient sample support to estimate the unknown covariances from the samples [88]. This is particularly an issue when performing multimodal data fusion, since the number of samples, *i.e.*, subjects, is typically much less than the dimension of the neurological data that is used. Thus, special attention must be paid both before performing an analysis, *i.e.*, when determining the similarity between datasets and their order—the dimension of the signal subspace—and while performing the analysis itself.

We define the similarity between two datasets as the number of common components that both datasets share, *i.e.*, those components that correlate across datasets, raising the issue of how to determine this number when the sample size is limited. One of the most popular exploratory techniques to estimate the number of common components between two datasets is based on the canonical correlation coefficients (CCCs) calculated using CCA [74] and defining a threshold for the level of the correlation, see *e.g.*, [89–93]. Other methods for estimating the number of common and distinctive sources include: orthogonal n -way partial least squares [94], generalized singular value decomposition [95], and distinctive and common components with simultaneous-component analysis [96]. These methods all assume sufficient sample support and thus perform poorly when the number of samples is not significantly greater than the number of observations. CCA, in particular, suffers greatly in the sample-poor regime, where all CCCs are significantly misestimated [97] and the highest CCCs, usually of greatest interest, may saturate at 1 [88], meaning that they provide no information about the true relationship between the datasets.

Since multimodal data is often quite noisy and of high dimensionality, dimension-

reduction using PCA is a crucial preprocessing step for avoiding the problem of over-fitting in subsequent analyses. However, the effectiveness of PCA is intimately tied to the problem of order-selection. For a single dataset, the most popular order-selection methods, such as AIC [39] or MDL [40], define the order based on ITC [43]. Though these methods have found widespread application in multimodal fusion, they are not directly applicable for two major reasons. The first is that almost all of these eigenvalue-based methods, with the notable exception of [98], assume sufficient sample support. If this is not true, such as for multimodal fusion using tCCA, where the number of subjects is much less than the dimension of the data, the performance of these methods deteriorates rapidly because the eigenvalues cannot be estimated accurately [98]. Additionally, these methods only report on the sources that have greatest energy in each dataset individually. Since we are interested in common components that are linked across datasets, the use of methods that focus solely on a single dataset is not a desirable solution to the question of order-selection for multimodal fusion. This provides the incentive to consider the problems of determining the degree of similarity and order *jointly*. Though not used in the context of medical imaging, there are methods that consider these two problems jointly, however these techniques: are heuristic [99, 100], will fail in the sample-poor regime [101], or will fail in the low SNR conditions inherent to multimodal fusion [102].

In this chapter, we discuss an effective method, PCA-CCA along with the order-selection rule from [97], for jointly determining the number of common sources across multimodal datasets and their order in the sample-poor regime and demonstrate its importance as a preliminary step for multimodal fusion. We first demonstrate the versatility and desirable performance of this technique through simulations. Then, we apply this new method to the pairwise combinations of fMRI, sMRI, and EEG data drawn from 14 patients with schizophrenia and 22 healthy controls while performing an AOD task and relate these

results to the pairwise fusion results obtained using tCCA. Through this application, we demonstrate a strong correlation between the number of common components estimated using PCA-CCA, *i.e.*, the similarity between datasets, and the number of statistically significant components estimated during the fusion analysis. This technique of investigating the pairwise combinations of datasets drawn from the same subjects provides a unique insight into the degree of complementarity between related data of different modalities.

3.2 Traditional and Sample-poor Hypothesis Test

Given M i.i.d. paired samples of $\mathbf{x}^{[1]} \in \mathbb{R}^{V_1}$ and $\mathbf{x}^{[2]} \in \mathbb{R}^{V_2}$ from the two-channel measurement model [97],

$$\begin{aligned}\mathbf{x}^{[1]} &= \mathbf{S}^{[1]}\mathbf{a}^{[1]} + \mathbf{n}^{[1]} \\ \mathbf{x}^{[2]} &= \mathbf{S}^{[2]}\mathbf{a}^{[2]} + \mathbf{n}^{[2]},\end{aligned}\tag{3.1}$$

where $\mathbf{a}^{[k]} \in \mathbb{R}^{d+f}$, $k = 1, 2$, are zero-mean jointly Gaussian random vectors with cross-covariance matrix, $\mathbf{R}_{\mathbf{a}^{[1]}\mathbf{a}^{[2]}} = E\{\mathbf{a}^{[1]}(\mathbf{a}^{[2]})^T\}$, given by

$$\mathbf{R}_{\mathbf{a}^{[1]}\mathbf{a}^{[2]}} = \begin{bmatrix} \mathbf{diag}(\rho_1\sigma_1^{[1]}\sigma_1^{[2]}, \dots, \rho_d\sigma_d^{[1]}\sigma_d^{[2]}) & \mathbf{0}_{d \times f} \\ \mathbf{0}_{f \times d} & \mathbf{0}_{f \times f} \end{bmatrix},$$

where $\sigma_i^{[k]}$ is the unknown standard deviation of signal component $a_i^{[k]}$ and ρ_i is the correlation coefficient between $a_i^{[1]}$ and $a_i^{[2]}$. Thus, both $\mathbf{a}^{[1]}$ and $\mathbf{a}^{[2]}$ have d correlated signals and f uncorrelated signals. Without loss of generality, we assume the auto-covariance matrices of $\mathbf{a}^{[1]}$ and $\mathbf{a}^{[2]}$ to be diagonal. The noise terms $\mathbf{n}^{[1]}$ and $\mathbf{n}^{[2]}$ are independent of each other, independent of the signals, and zero-mean Gaussian with unknown covariance matrices. Additionally, without loss of generality we assume that $\mathbf{S}^{[1]}$ and $\mathbf{S}^{[2]}$ are of full column rank

and, like the dimensions d and f , are fixed but unknown.

We collect the M sample pairs into data matrices $\mathbf{X}^{[1]} = [\mathbf{x}^{[1]}(1), \dots, \mathbf{x}^{[1]}(M)]$ and $\mathbf{X}^{[2]} = [\mathbf{x}^{[2]}(1), \dots, \mathbf{x}^{[2]}(M)]$. When performing CCA in the where $M < V_1 + V_2$, at least $V_1 + V_2 - M$ of the sample canonical correlation coefficients, $\hat{k}_i, i = 1, \dots, q, q = \min(V_1, V_2)$, will be identically 1 regardless of the values of ρ_i and thus do not provide any information about the relationship between $\mathbf{a}^{[1]}$ and $\mathbf{a}^{[2]}$ [88]. Moreover, even in the case where M is greater, but not significantly greater, than $V_1 + V_2$, the sample canonical correlations may significantly overestimate the population canonical correlations [97]. This result provides the incentive to estimate a suitable rank, C , in order to reduce the dimensions of $\mathbf{X}^{[1]}$ and $\mathbf{X}^{[2]}$, thus allowing accurate estimation of the number of correlated signals.

A classical way of estimating d is by assuming that the sources are drawn from a multivariate Gaussian distribution and applying a sequence of binary hypothesis tests [103, 104]. The test begins with $s = 0$ and compares the two hypotheses $H_0 : d = s$ and $H_1 : d > s$. If the null hypothesis is rejected, then s is increased by one and the test is repeated, until either the null hypothesis is not rejected or $s = q$. This test is based on the Bartlett-Lawley statistic [103, 104], which is given by

$$L(s) = \left(M - s - \frac{V_1 + V_2 + 1}{2} + \sum_{i=1}^s \hat{k}_i^{-2} \right) \log \prod_{i=s+1}^q (1 - \hat{k}_i^2), \quad (3.2)$$

and is asymptotically distributed under H_0 as χ^2 with $(V_1 - s)(V_2 - s)$ degrees of freedom. The fact that the test statistic is distributed according to the χ^2 distribution enables the determination of a threshold, $T(s)$, to meet a given probability of false alarm, P_{FA} , for the test. A major constraint of the traditional framework is the assumption of sufficient samples, *i.e.*, that the \hat{k}_i 's are accurate estimates of the true k_i 's, making it inapplicable for the sample-poor regime.

As proposed in [97], the sample-poor version of the classical hypothesis test selects

$$\hat{d} = \max_{r \in \mathcal{R}} \min_{s \in \mathcal{S}} \{s : L(s, c) < T(s, c)\}, \quad (3.3)$$

where $\mathcal{R} = \{1, \dots, r_{\max}\}$, $\mathcal{S} = \{0, \dots, r - 1\}$, and $r_{\max} = \min(M/2, q)$ and the test statistic is

$$J(s, r) = \left(M - s - r - \frac{1}{2} + \sum_{i=1}^s \hat{k}_i^{-2}(r) \right) \log \prod_{i=s+1}^r (1 - \hat{k}_i^2(r)).$$

Note in the above expression that the estimated canonical correlations are functions of the reduced rank, r ; if r is too big, then \hat{k}_i will be over-estimated, while if it is too small, then \hat{k}_i will be under-estimated [97].

By limiting the dimension through r_{\max} , we can ensure that none of the estimated canonical correlations will saturate at 1. Note that to guarantee that the rank-reduced statistic under H_0 is still asymptotically distributed as χ^2 with $(r - s)^2$ degrees of freedom, we must pick r_{\max} “sufficiently” smaller than $M/2$. This is critical in order to compute the test threshold, $T(s, r)$, for a given probability of false alarm. Based on our numerical simulations, a good value for r_{\max} seems to be $\lfloor M/3 \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function [97]. However, note that there is a certain degree of robustness of the results to the selection of the value of r_{\max} observed in real data. The value for r that leads to \hat{d} in (3.3) is chosen as the PCA rank.

3.2.1 Determining the Probability of False Alarm using ITC

A fundamental issue with hypothesis tests is the need to set a probability of false alarm, which determines the test threshold. However, the ideal value for this probability of false alarm is application-dependent and can be difficult to determine for real data. This provides the incentive to use a threshold that is based on ITC. The general MDL expression for detecting the number of correlated components in a two channel model is given by [105]

$$I_{\text{MDL}}(s) = M \log \left(\prod_{i=1}^s (1 - \hat{k}_i^2) \right) + \log(M)s(m + n - s), \quad (3.4)$$

where the number of correlated components is determined by the value of s which minimizes (3.4).

We can modify (3.4) in a similar manner as (3.2) to account for the PCA step, resulting in the final MDL criterion [97]

$$I_{\text{MDL}}(s, r) = M \log \left(\prod_{i=1}^s (1 - \hat{k}_i^2(r)) \right) + \log(M)s(2r - s). \quad (3.5)$$

The decision rule (3.3) is thus replaced with

$$\hat{d} = \max_{r \in \mathcal{R}} \operatorname{argmin}_{s \in \mathcal{S}} I_{\text{MDL}}(r, s), \quad (3.6)$$

and, because of its desirable performance, as we demonstrate through simulations in the next sections, is used to compute the number of common components for the multimodal data considered in this chapter.

3.3 Simulations

3.3.1 Variable Per-Component SNR Simulation Setup

For real brain data, the ground truth, *i.e.*, the true latent sources and subject covariances, is not available. Thus, in order to test the performance of PCA-CCA, we generate simulated data according to (3.1) and examine the probability of correctly estimating the true number of correlated components as the dimension of the datasets, V_1 and V_2 , increases from 10 to 10,000. We compare the performance of PCA-CCA to that of two classical ITC-based methods, AIC and MDL. These methods use the AIC and MDL framework in [43] in the PCA step to select r , and then use the AIC and MDL framework described in [105] to estimate the common order. For these simulations, we have $M = 50$ samples and $d = 2$

pairs of correlated components each with variance of 2 and correlation coefficients, $\rho_1 = 0.8$ and $\rho_2 = 0.7$. The number of uncorrelated components in $\mathbf{a}^{[1]}$ and $\mathbf{a}^{[2]}$, f , is 3. The variances of the uncorrelated signals and the white noise are both 1. We select these parameters in order to align with properties of real neurological data. The entries of the mixing matrices $\mathbf{S}^{[k]}$ are independently drawn from uniformly distributed random numbers between 0 and 1. We consider two values, 0.001 and 0.005, for probability of false alarm, P_{FA} . We also use the MDL-based method described previously, which does not require the selection of a probability of false alarm. The plots that we present here are the average result of 1,000 independent Monte-Carlo simulations, where, for each simulation, we apply every method to the same data.

We repeat this simulation three times, changing only the distribution from which the sources, $\mathbf{a}^{[k]}$, are drawn. In the first case, we consider sources drawn from a zero mean Gaussian distribution, which aligns with the form typically used in all derivations [42, 43, 97, 103, 104]. The subsequent case considers sources that are uniformly distributed with zero mean and the same variance as the previous case. Such sub-Gaussian or platykurtic sources are a good approximation of the subject covariations, since if there is a difference between the two groups then the distribution describing the subject covariations will be bi-modal and hence sub-Gaussian. Thus, the performance of PCA-CCA in this case is of great interest. The final case considered is where the sources are Laplacian distributed with zero mean and the same variance as the previous distributions. Due to the ability of the Laplacian distribution to provide good approximations of empirical fMRI source distributions [65, 106], studying the performance of PCA-CCA in this regime is of interest for the fusion of two fMRI datasets.

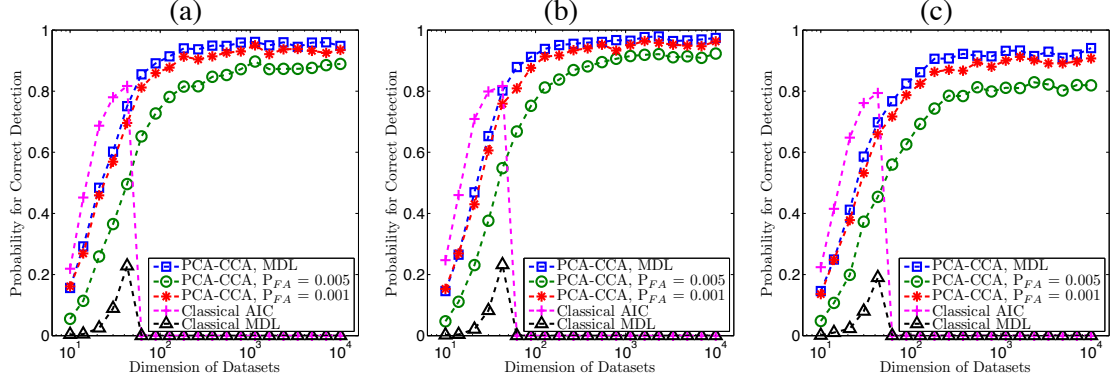


FIG. 3.1. Probability of correctly detecting the number of common components versus the dimension of the datasets, V_1 and V_2 , when the latent sources are (a) Gaussian, (b) uniformly, and (c) Laplacian distributed. Note that for these simulations, $V_1 = V_2$.

3.3.2 Variable Per-Component SNR Simulation Results

Figure 3.1 shows the high level of performance obtained using PCA-CCA together with the order-selection rule in [97] for the three source distributions considered. The results indicate that PCA-CCA is robust to model mismatch in the latent source distribution. In all cases, the performance of PCA-CCA improves as the dimensionality of the datasets increases. The reason for this is that as the dimension of the datasets increases, the signal eigenvalues of $\mathbf{R}_{\mathbf{x}^{[i]}\mathbf{x}^{[i]}} = \mathbf{S}^{[i]}\mathbf{R}_{\mathbf{a}^{[i]}\mathbf{a}^{[i]}}(\mathbf{S}^{[i]})^T + \sigma_n^2\mathbf{I}_M$, $i = 1, 2$, where $\mathbf{R}_{\mathbf{a}^{[i]}\mathbf{a}^{[i]}}$ is the population covariance matrix of the correlated and uncorrelated signals, $M = V_1 = V_2$ is the dimension of the system, σ_n^2 is the variance of the noise, and \mathbf{I}_N is a M -dimensional identity matrix, increase, whereas the noise eigenvalues remain unchanged. Thus, the per-component SNR, *i.e.*, the ratio of signal eigenvalue to the noise eigenvalue, increases as the dimension increases, thus improving the probability of detection. We see that the performance of the classical methods, AIC and MDL, deteriorates rapidly as the datasets become increasingly sample-poor because, in this regime, the eigenvalues cannot be estimated accurately. Finally, note that the MDL-based method consistently has the highest performance, indicating

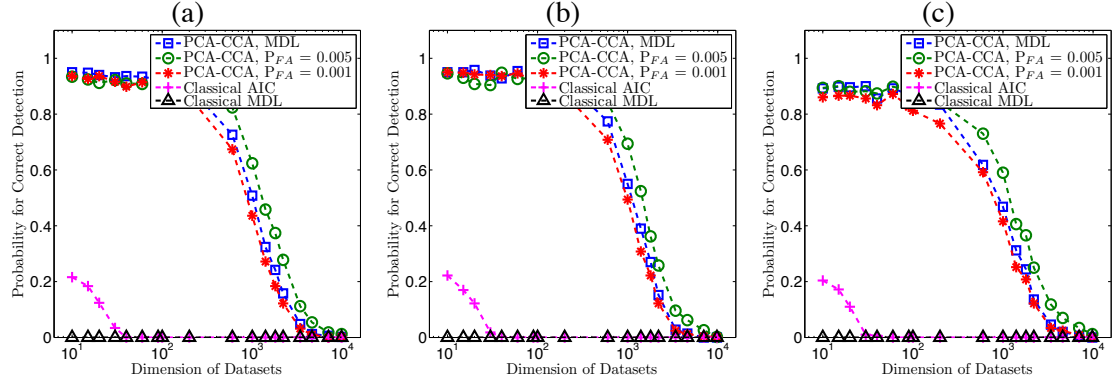


FIG. 3.2. Probability of correctly detecting the number of common components versus the dimension of the datasets, V_1 and V_2 , when the latent sources are (a) Gaussian, (b) uniformly, and (c) Laplacian distributed. Note that for these simulations, $V_1 = V_2$.

the benefits of its use over that of an arbitrary selection of a probability of false alarm.

3.3.3 Fixed Per-Component SNR Simulation Setup

In order to investigate the case where the per-component SNR does not increase with the dimension, we simulate the case where the signal eigenvalues remain constant as the dimension of the data increases and display the results below. We design the simulation using similar parameters to the previous simulation, with two major exceptions: first, the number of independent Monte-Carlo simulations is 500 and second, the columns of $\mathbf{S}^{[k]}$, $k = 1, 2$ are unitary and not independently drawn from uniformly distributed random numbers between 0 and 1. This second change ensures that the signal eigenvalues do not change as we modify the dimension, V_1, V_2 . Also, in order to consider a less noisy case than in the previous simulation, for this simulation the variance of the uncorrelated signals is 3 and the variance of the correlated signals is 10.

3.3.4 Fixed Per-Component SNR Simulation Results

From Figure 3.2, we see that when the signal eigenvalues do not change as the dimension increases, the detection performance of all detectors decreases. The reason for this is due to the existence of an asymptotic threshold below which signal eigenvalues are classified as noise eigenvalues, leading to a drop in the estimated number of common components as seen in equation 18 in [98]. This threshold depends on the ratio of dimension of the datasets, V_1, V_2 , to M . Thus, if the dimension of the datasets increases, the ratio will increase and the signal eigenvalues will be classified as noise eigenvalues. Note that even in this case, PCA-CCA still significantly outperforms the traditional methods.

3.4 Fusion Results Using Real Multimodal Data

We depict the fusion results using tCCA in Figure 3.3. Note that for these results, 15 components were estimated, which corresponds to the mean order found using PCA-CCA on the pairwise combination of datasets, which we show in Table 3.1. The components displayed are those that were found to be both biologically meaningful and statistically significant, *i.e.*, they had activations corresponding to biologically meaningful areas and p -values in terms of group differences smaller than 0.05. Biological meaning is assessed based on prior information about the spatial and temporal properties of the components. For the fMRI and sMRI components, smooth and focal regions are expected, while ERP components should display a smooth response with peaks corresponding to those in the group response. When using the fMRI and sMRI data together, one component is found. This component corresponds to increased frontal lobe activation in the fMRI section and an increase in gray matter concentrations in the motor cortex in the sMRI of controls versus patients. When the fMRI and ERP data are used in the analysis, a single one component is found that is both statistically significant and physically meaningful. This component

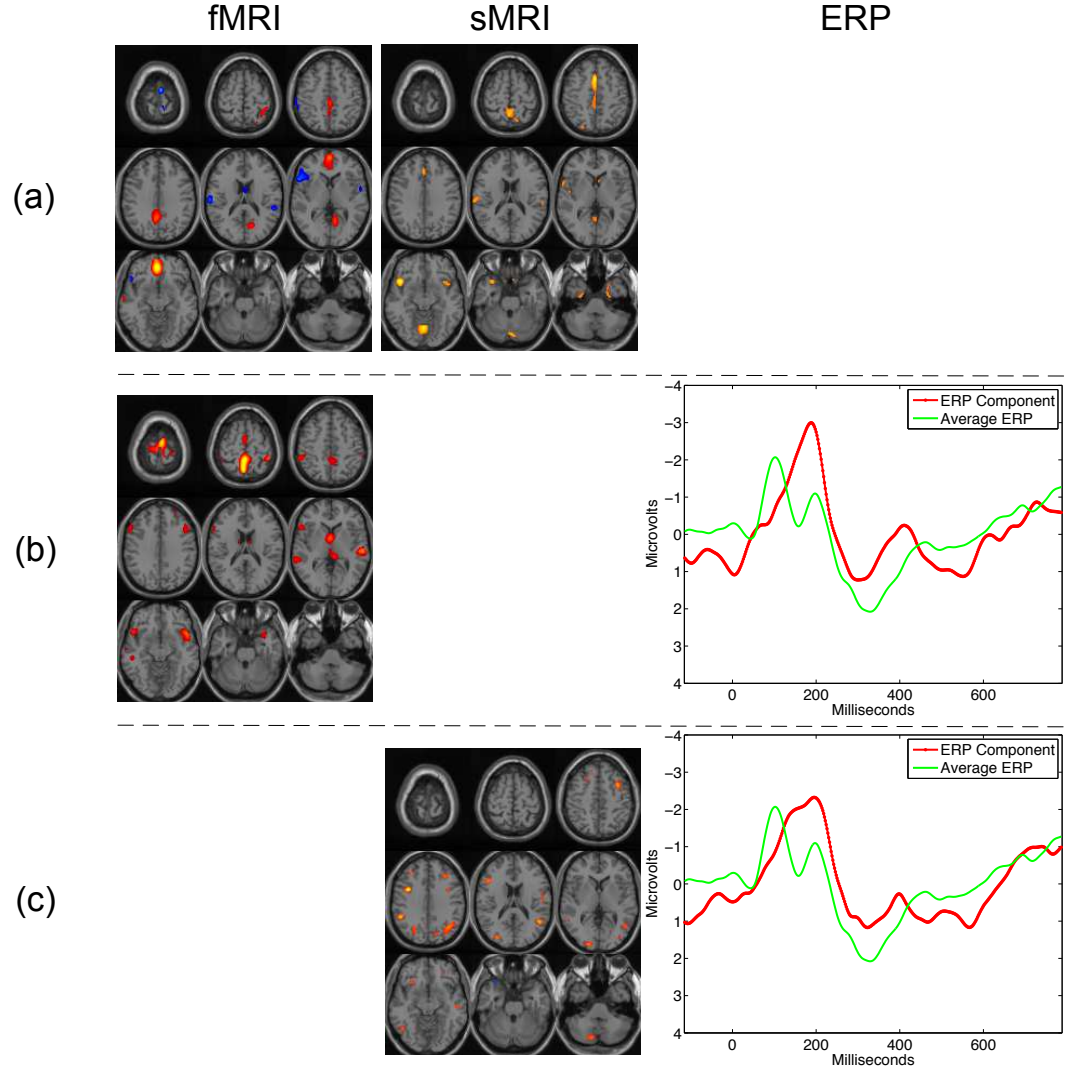


FIG. 3.3. Meaningful and statistically significant components generated using tCCA with pairwise combinations of the three modalities. The uncorrected significance (p -values) for the components are (a) 0.027 for the fMRI-sMRI component, (b) 0.003 for the fMRI-ERP component, and (c) 0.018 for the sMRI-ERP component. All spatial maps are Z-maps threshold at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients.

Table 3.1. Number of significant components at two p -values, PCA-CCA estimated common components, d , and estimated order, C , for the three pairwise combinations of modalities. We assess statistical significance using a two-sample t -test run on the rows of the estimated subject covariations, $\hat{\mathbf{A}}^{[k]}$, $k = 1, 2$, calculated using tCCA.

	fMRI-sMRI	fMRI-ERP	sMRI-ERP
Number of significant components, $p = 0.05$	2	1	1
Number of significant components, $p = 0.1$	5	4	2
Number of common components using PCA-CCA	4	3	2
Estimated order using PCA-CCA	17	17	12

corresponds to increased sensory-motor and auditory activation in the fMRI component and increased activation in the N2\P3 complex in the ERP of controls versus patients. These regions correspond to areas that have been previously noted as affected in schizophrenia. We see a similar N2\P3 complex in the ERP component generated using the sMRI and ERP data. Note that, as we describe in Chapter 5, these discriminative components can be leveraged to classify a new subject as either a patient with schizophrenia or a healthy control [72].

Using PCA-CCA and the MDL criterion that we describe in Section 3.2.1, we can estimate the number of common components for each of the pairwise combinations of the fMRI, sMRI, and ERP data. This estimation of the number of common components allows us to measure the degree of similarity between the datasets in a meaningful way, by defining greater similarity in terms of a greater number of common components. The results of this analysis are displayed in Table 3.1, where we also show the number of statistically significant components found for each combination of modalities. We declare components significant based upon the result of a two-sample t -test run on the rows of the estimated subject covariations, $\hat{\mathbf{A}}^{[k]}$, $k = 1, 2$, calculated using tCCA. We see that the fMRI and sMRI data have the greatest degree of similarity, since they share the greatest number of common components estimated using PCA-CCA. This result makes sense since both modalities are MRI data of spatial nature and thus they would be expected to have a greater degree of

similarity. The result also corresponds to the number of significant components that we find after performing fusion using tCCA. Since the fMRI-sMRI combination has the greatest number of significant components, there is a greater number of chances of finding common significant components. This explains why we observe that greatest number of meaningful components using the combination of fMRI and sMRI. Using PCA-CCA, we find that the combination of fMRI and ERP data produce the second greatest number of common components. This result again makes sense, since both the fMRI and ERP data describe functional changes in the brain, so they exhibit a fairly large degree of similarity. The result correlates to the results that we obtain in terms of number of significant components found after performing fusion using tCCA on the modality pair of fMRI-ERP. From the results obtained using PCA-CCA, the least similar modalities are the sMRI and ERP, since the sMRI-ERP combination produced the fewest number of common components. This result makes sense since the sMRI data is structural whereas the ERP data is functional. Thus, using PCA-CCA, we can measure the degree of similarity between datasets and the results can be predictive of the fusion results using tCCA.

3.5 Discussion

The successful application of PCA-CCA to determine the degree of similarity between pairwise combinations of multimodal datasets naturally leads to the issue of extensions beyond two datasets. As we discuss in Section 2.3.2, for fusion using tCCA, the straightforward extension is the use of tMCCA [75], using one of the five cost functions. We display the results of the application of tMCCA using the GENVAR cost, to the fMRI, sMRI, and ERP data in this study in Figure 3.4. As we expect, we can see a great degree of correlation between the tCCA results and those achieved using tMCCA. The motor, auditory, and visual activation in the second component bear striking resemblance to those of the fMRI-

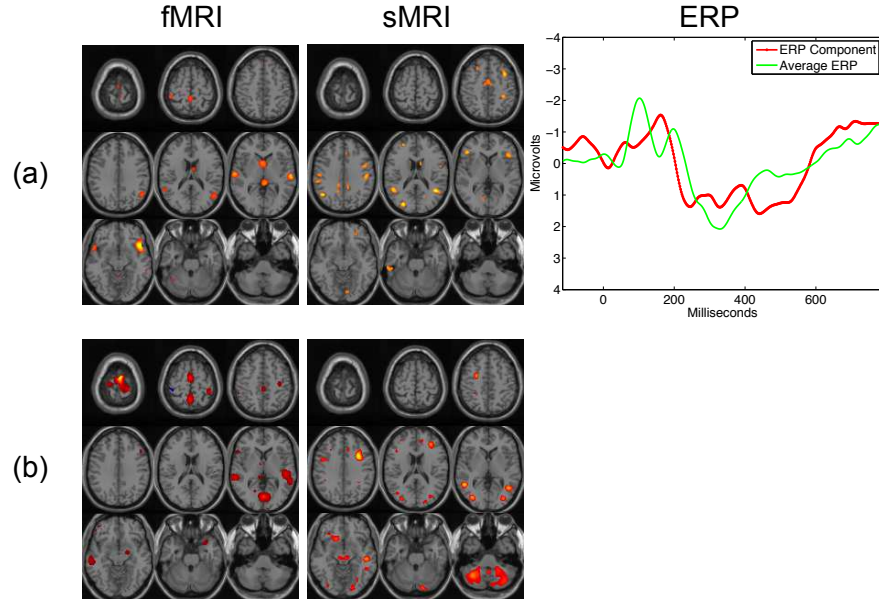


FIG. 3.4. Meaningful and statistically significant components generated using MCCA with the GENVAR cost. The uncorrected significance (p -values) for the components are (a) 0.014 for the first component and (b) 0.018 for the second component. All spatial maps are Z-maps thresholded at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients.

ERP component. In addition, the N2\P3 activation that we observe in the ERP components that are generated using tCCA is similar to that of the ERP component generated using tMCCA.

However, a similar extension of PCA-CCA to more than two datasets remains an open problem. This is due to the fact that in the case of two datasets the number of non-zero canonical correlations defines the degree of correlation between the two sets. However, for more than two datasets, the canonical correlations cannot be used to define the number of correlated components. To understand the reason for this, consider the case where we have three datasets with four independent signals in each. Only the first two sources in dataset 1 are correlated with the first two sources in dataset 2. Similarly, only the first and third sources in dataset 2 are correlated with the first and third sources in dataset 3. Finally, only

the first and fourth sources in dataset 1 are correlated with the first and fourth sources in dataset 3. In this example, there are two correlated signals in each pair of datasets but there is just one signal, *i.e.*, the first signal in each dataset that is correlated across all the datasets. In order to identify this common signal, we expect a non-zero correlation between all three pairwise combinations of the first signal components of the three datasets only. However, tMCCA would generate non-zero correlations for all pairwise combinations of the four signals from all datasets, regardless of the cost used, indicating that all of the components are common to all datasets, which is not the case.

Though a true multiset extension of PCA-CCA is preferable, an interim solution is the following heuristic method, which can estimate the number of correlated components among all three datasets. Note that this method can be extended to an arbitrary number of datasets, though it is only presented for three datasets here. Consider three datasets generated in the same manner as in (3.1)

$$\begin{aligned}\mathbf{x}^{[1]} &= \mathbf{S}^{[1]}\mathbf{a}^{[1]} + \mathbf{n}^{[1]}, \\ \mathbf{x}^{[2]} &= \mathbf{S}^{[2]}\mathbf{a}^{[2]} + \mathbf{n}^{[2]}, \\ \mathbf{x}^{[3]} &= \mathbf{S}^{[3]}\mathbf{a}^{[3]} + \mathbf{n}^{[3]}.\end{aligned}$$

For datasets $\mathbf{x}^{[1]}$ and $\mathbf{x}^{[2]}$, by applying the PCA-CCA approach, we can estimate the number of correlated signals between the first and second datasets, $\hat{d}^{[1,2]}$, and the optimal rank, $C^{[1,2]}$, that PCA should keep for $\mathbf{x}^{[1]}$ and $\mathbf{x}^{[2]}$. With these estimates, the canonical sources, $\omega_1^{[1,2]}$ and $\omega_2^{[1,2]}$, associated with the $\hat{d}^{[1,2]}$ largest canonical correlations may be obtained from the $C^{[1,2]}$ -approximation of $\mathbf{x}^{[1]}$ and $\mathbf{x}^{[2]}$, respectively. The above procedure for $\mathbf{x}^{[1]}$ and $\mathbf{x}^{[2]}$ can be analogously applied for the pairwise combinations $\mathbf{x}^{[2]}$ and $\mathbf{x}^{[3]}$ as well as $\mathbf{x}^{[1]}$ and $\mathbf{x}^{[3]}$. Due to the fact that the canonical source, *e.g.*, some row in $\omega_1^{[1,2]}$, is equal to the corresponding original source, *e.g.*, some row in $\mathbf{a}^{[1]}$, up to a scaling factor, we can find three

canonical sources, $\omega_1^{[1,2]}, \omega_2^{[2,3]}, \omega_3^{[1,3]}$ whose cross-correlation matrices, $E \left\{ \omega_1^{[1,2]} \left(\omega_2^{[2,3]} \right)^T \right\}$, $E \left\{ \omega_1^{[1,2]} \left(\omega_3^{[1,3]} \right)^T \right\}$, $E \left\{ \omega_2^{[2,3]} \left(\omega_3^{[1,3]} \right)^T \right\}$, are all of rank d , where d is the number of common signals among three datasets. Then, the ITC-based detection scheme described in [107] can be used. Applying this method to the fMRI, sMRI, and ERP data, results in the identification of one common signal. The result using this method corresponds to what we observe with tMCCA, where there is only one significant component that is common to all three modalities.

3.6 Summary

The collection of multimodal data is becoming increasingly common in medical studies, raising the issue of how best to combine data that are, fundamentally, different. In this chapter, we propose the use of PCA-CCA and the order-selection rule of [97] to determine the number of correlated signals, degree of similarity, between multimodal neurological data in the sample-poor regime. The results using PCA-CCA indicate that this method can not only measure the degree of similarity between dissimilar datasets but also provides an accurate measure of prediction of the significance of components generated by tCCA and tMCCA. This predictive capability of PCA-CCA enables the determination of meaningful combinations of modalities prior to a fusion analysis.

Chapter 4

FUSION OF THREE MULTIMODAL NEUROIMAGING DATASETS

Despite the benefits of fusing multiple modalities, generally only two datasets are combined. However, the fusion of three or more modalities provides a unique opportunity to study what each modality contributes to the fusion result and which combinations of modalities are best to fuse. In this chapter, we compare the performance of the same techniques that we discuss in Chapter 2 on simulated neuroimaging data in order to test the robustness of the techniques to deviations from their assumptions. We also apply the methods that we develop in the previous chapter to fuse fMRI, sMRI, and ERP data drawn from patients with schizophrenia and healthy controls. We find that the number of subjects and the connections between the modalities can greatly affect the results of the different fusion methods. We also find that by comparing different methods on different combinations of the datasets, we can gain greater understanding about the connections between the modalities. Our results show that when the modalities share a large amount of discriminatory power, such as for the combination of fMRI and ERP, jICA, with its more constrained model may be preferable, since it is more robust to noise. However, when this link between the datasets is weak, such as for the combinations of fMRI and sMRI or sMRI and ERP, the more flexible tCCA or tIVA models may be more reasonable.

4.1 Introduction

The fusion of data from multiple sensors can enable greater understanding of the problem under study than analyzing the datasets individually [10]. Despite this fact, nearly every study that fuses multimodality neuroimaging datasets from the same subjects only considers two modalities, with a few exceptions, see *e.g.*, [12, 28, 38, 108–113]. Much in the same way that the fusion of two modalities can improve upon separate analyses of the modalities individually, the fusion of multiple modalities has the potential to improve the information gained and inferences further. However, there are a number of practical considerations that must be accounted for before applying a method to a particular dataset, such as: the number of subjects and whether the modeling assumptions of the methods are realistic for the data.

In order to explore these issues, we adopt a two-part investigation. In the first part, we use simulated data to explore the effects of model mismatch and noise on the results of different data fusion methods. Such an investigation is important because, since the ground truth is known, it grants us the ability to objectively compare the results of different fusion methods. In the second part of the investigation, we apply the data fusion methods to real fMRI, sMRI, and ERP data drawn from 36 subjects, 14 subjects with schizophrenia and 22 healthy controls. Note that since this multimodality data is from relatively few subjects, we only compare the methods jICA, tCCA, tMCCA using different costs, and tIVA-G. This investigation allows us the ability to see how the modeling assumptions of each method determine their results and how much each dataset contributes to the fusion analysis by comparing the results when it is included to the results when it is not.

4.2 Multimodal Data

The data that we use in this chapter is fMRI, sMRI, and ERP data drawn from 36 subjects, 14 patients with schizophrenia and 22 healthy controls, resulting in datasets of dimensions $\mathbf{X}_{\text{fMRI}} \in \mathbb{R}^{36 \times 60,261}$, $\mathbf{X}_{\text{sMRI}} \in \mathbb{R}^{36 \times 306,626}$, and $\mathbf{X}_{\text{ERP}} \in \mathbb{R}^{36 \times 451}$. Due to the issues associated with jICA that we discuss in Chapter 2, namely the fact that jICA inherently weighs each modality by its dimensionality, we concatenate the ERP with itself 100 times in order to make it of approximately the same dimensionality as the fMRI and sMRI. Note that, in order to facilitate comparisons between methods, we estimate the same number of components, 15, for each method for all combinations of datasets. We select this number because it corresponds to the mean of the orders that we estimate using PCA-CCA on the pairwise combinations of the modalities. Note that before displaying the statistically significant components, we evaluate the components using prior knowledge about the properties of the imaging modalities to identify whether or not the components are physically meaningful. For the fMRI, we expect smooth and focal regions of activation/deactivation. Similarly, for the sMRI, we expect smooth a focal regions within the gray matter. The physically meaningful ERP components correspond to smooth responses, generally aligning with peaks in the subject-averaged ERP.

4.3 Simulation

4.3.1 Simulation Setup

In order to create the simulated data, we generate ten sources, each with 500 i.i.d. samples drawn from a zero-mean Laplacian distribution with a standard deviation of 4. We linearly mix these sources using separate mixing matrices of dimensions 50×10 . We model the columns of each $\mathbf{A}^{[k]}$ in the following manner: $\mathbf{a}_1^{[k]} = \mathbf{g} + \mathbf{n}_1^{[k]}$ and $\mathbf{a}_i^{[k]} = \mathbf{n}_i^{[k]}$, $i = 2, \dots, 10$, where \mathbf{g} is a step of height 1.5 with a length of 25 and $\mathbf{n}_i^{[k]}$ are 50 i.i.d. samples from a zero-

mean and unit variance Gaussian distribution. Thus, the first column simulates a difference between two groups, each of size 25 subjects. Note that we adjust the standard deviations of the noise for the columns not representing the group difference, *i.e.*, $\sigma_{\mathbf{n}_i}, i = 2, \dots, 10$ so that they have the same approximate value as the standard deviation of $\mathbf{a}_1^{[k]}$. The column of the mixing matrix with the step-type response, establishes the connection between the datasets, *i.e.*, results in components that have a common group difference. The number of samples in $\mathbf{a}_i^{[k]}$, representing subjects, is small, meaning that we consider a case similar to that of real neuroimaging data. For all of the simulations in this chapter, we consider two scenarios, the first, whose results are shown in Figure 4.1, where there are only two datasets and the second, whose results are shown in Figure 4.2, where there are three. For the case where there are two datasets, we compare jICA using the EBM algorithm with tCCA and tIVA-G. For the case where there are three datasets, we compare jICA using the EBM algorithm with tIVA-G and tMCCA using the MAXVAR and GENVAR cost functions. We select these algorithms due to the fact that the number of subjects is quite limited, thus more complicated IVA algorithms cannot be used for tIVA. Note that the identifiability condition of tIVA-G is only satisfied for \mathbf{a}_1 , defined as $[\mathbf{a}_1^{[1]} \mathbf{a}_1^{[2]}]^T$ or $[\mathbf{a}_1^{[1]} \mathbf{a}_1^{[2]} \mathbf{a}_1^{[3]}]^T$ for the two or three dataset cases, respectively [84]. For jICA, we horizontally concatenate the datasets and reduce the dimension to 10, thus resulting in a joint dataset that is 10×1000 for the two dataset scenario and a joint dataset that is 10×1500 for the three dataset scenario. For tCCA/tMCCA/tIVA-G, we transpose each individual dataset and then reduce the dimension to 10, resulting in matrices that are of dimension 10×50 . The results that we show in Figures 4.1 and 4.2 are the results of 100 independent Monte-Carlo simulations, where we show the average correlation of the estimated component whose estimated subject covariation had the highest average t -statistic and the original component, which had the step-type profile.

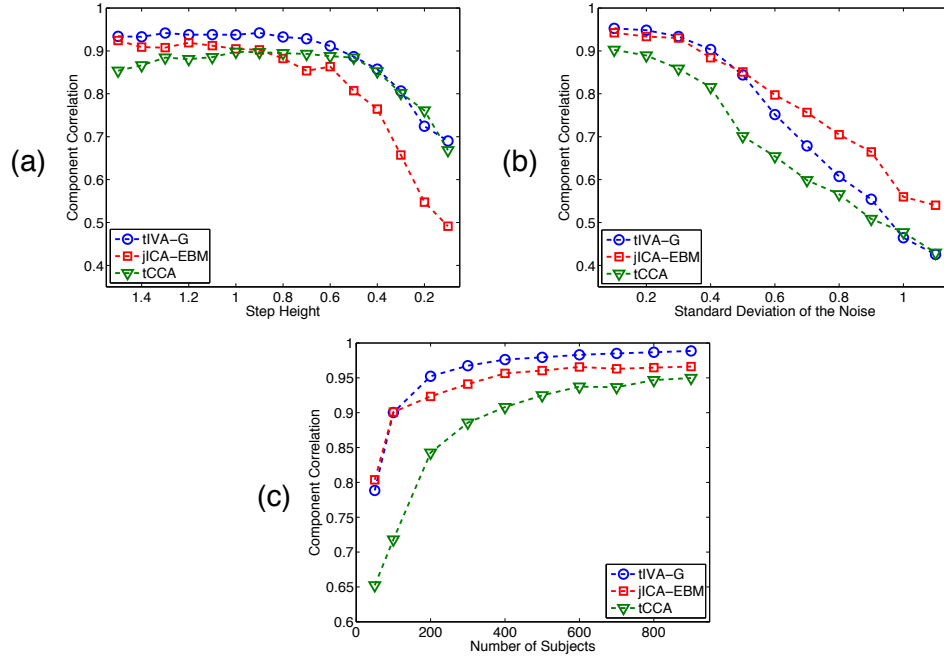


FIG. 4.1. Estimation performance for the common component with two datasets as (a) the height of the step in $\mathbf{a}_1^{[1]}$ decreases (b) $\mathbf{n}_1^{[1]}$ increases, and (c) the number of subjects increases.

4.3.2 Simulation Results and Discussion

In the two dataset scenario, shown in Figure 4.1(a) and (b), we control the link between the two datasets by either changing the height of the step or by keeping the height constant and changing $\mathbf{n}_1^{[1]}$. Thus, we study the performance of the two models when the underlying common structure between the two datasets becomes weaker as well as when it is strong but the level of noise is high. Note that, for the results shown in Figure 4.1, the value of $\sigma_{\mathbf{n}_1}$ is 0.3. We see in Figure 4.1(a) that when the connection becomes weaker due to the change in the step height, tIVA-G provides better performance than jICA, since the two mixing matrices become less similar, thus violating a central assumption of jICA. However, we see in Figure 4.1(b) that jICA performs better as the noise level increases, since the underlying connection between the two datasets for the discriminative component is strong implying that jICA tends to “average out” the effects of additive noise. We should note

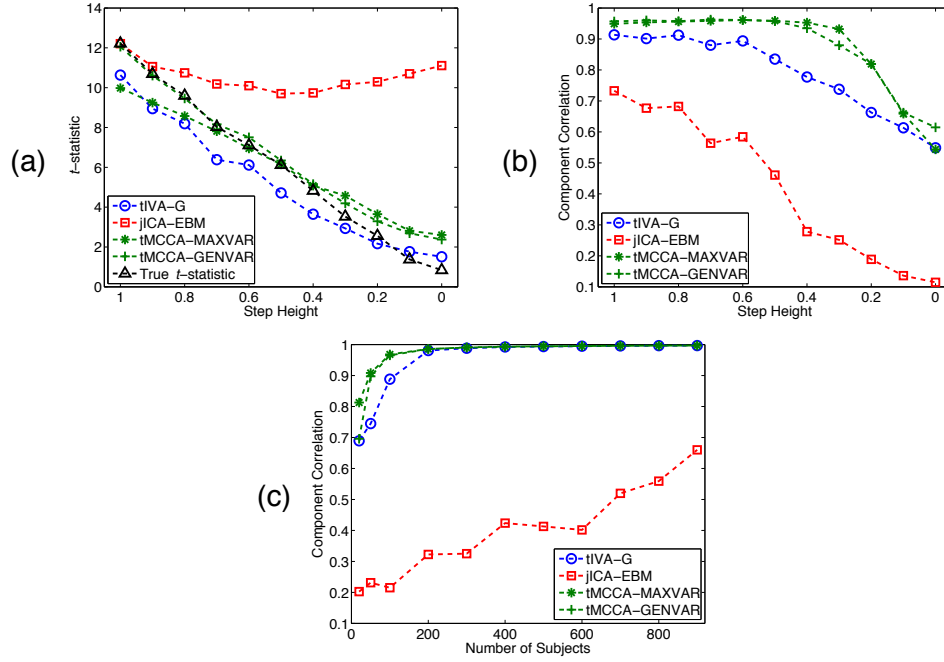


FIG. 4.2. Estimation performance for the common component with three datasets as the height of the step decreases in terms of (a) t -statistic for the most significant component, (b) and (c) correlation of the most significant component to the original source.

that tCCA has worse performance than tIVA-G in this case, likely due to the orthogonality constraint for CCA, which limits the search space for the optimal demixing matrices. Since the mixing matrices become more orthogonal as the height of the step decreases, the relative performance of tCCA compared with tIVA-G increases. Figure 4.1(c) shows the change in performance of the data fusion techniques as the number of subjects increases for fixed step height of 1.5 and $\mathbf{n}_1^{[1]} = 0.3$, demonstrating the advantages for tIVA model as the number of subjects increases.

The second set of simulations considers three datasets, again with one common component across all three. We first study the effect of decreasing correlation for one dataset by decreasing the step height for this one component, while the step for the other two are kept constant. In Figure 4.2(a), we show trends in the t -statistic and the estimation performance, which we measure using the correlation with the estimated source to the original for the

components for which the step size is altered. Note that the significance for the component estimated by jICA tends to increase for step heights below 0.5, while the estimation of the component itself, *i.e.*, its correlation with the original, starts to significantly decrease during the same interval, as shown in Figure 4.2(b). On the other hand, tMCCA using both cost functions and tIVA-G yield reliable estimates for the t -statistics and correlation values, as seen in Figure 4.2(a) and (b). We can see in Figure 4.2(b) that tMCCA-MAXVAR and tMMCA-GENVAR are slightly better than tIVA-G. We can attribute this to the fact that the mixing matrices in this scenario are closer to being orthogonal, especially as the height of the step function decreases, thus favoring the simpler methods. In the results shown in Figure 4.2(c), we keep the step height at 0.3, while the other two had a value of 1. We see that as the number of subjects increases, the performance of all methods increases, though jICA has the worst performance due the violation of the identical mixing matrix assumption.

4.4 Multimodal Fusion of FMRI, SMRI, and ERP

The fusion results using all three modalities for jICA, tMCCA using MAXVAR and GENVAR, and tIVA-G are shown in Figures 4.3 and 4.5, respectively. We first describe the jICA results and then explain their connection with the tMCCA and tIVA-G results.

4.4.1 JICA Results and Discussion

Figure 4.3 displays the results of fusing the three modalities of fMRI, sMRI, and ERP using jICA. The component shown in Figure 4.3(a) identifies differences between patients and controls in the anterior motor as well as temporal regions in the fMRI and is associated with a decrease in gray matter concentration for patients over controls in the frontal lobe in the sMRI and the P2/N2 transition in the ERP. The second component, shown in Figure 4.3(b), has differences between patients and controls in the sensorimotor as well as temporal regions in the fMRI and is associated with increases in gray matter concentration for

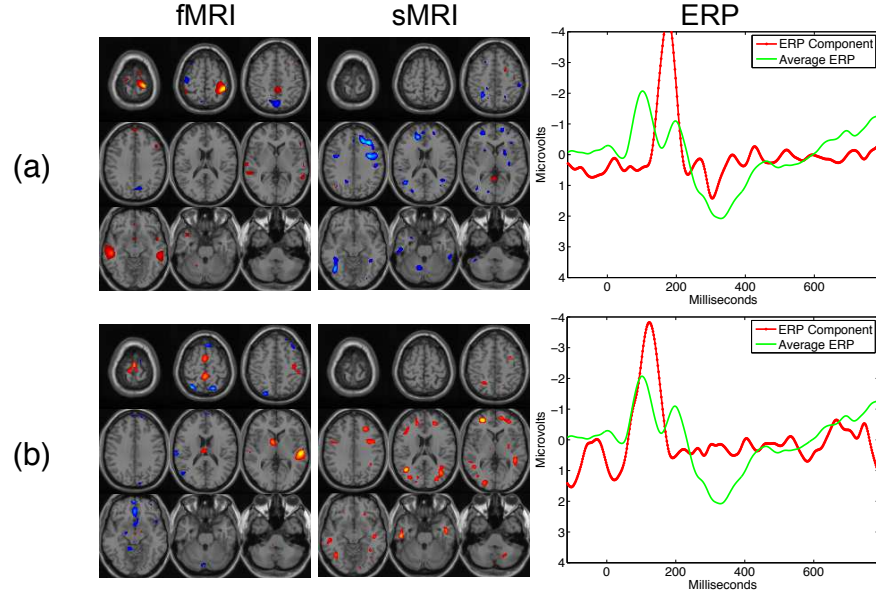


FIG. 4.3. Significant and biologically meaningful components generated for fusion of all three datasets using jICA using the EBM algorithm. The uncorrected significance values are (a) 0.0042 and (b) 0.0097. All spatial maps are Z-maps thresholded at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients.

controls over patients throughout the frontal and occipital lobes in the sMRI and the N1/P2 transition in the ERP. In order to probe the contributions of the modalities to the final result, we also perform pairwise fusion analyses using the combination of fMRI and ERP as well as sMRI and ERP. We display the results in Figure 4.4.

We note that the significant components found using only the fMRI and ERP modalities, displayed in Figure 4.4 (a) and (b), are nearly the same as the components found using all three modalities, implying that the fMRI and ERP datasets are the primary ones driving the result. However, we should note two things: first, the component found when only fusing the sMRI and ERP is fairly similar to the component displayed in Figure 4.3(a), implying that the sMRI is providing some discriminatory information. Also, we note that the rapid transitions between the negative and positive peaks in Figure 4.3(a) and (c) is hard to interpret physically. Thus, in this application, we note an increase in the “meaningful-

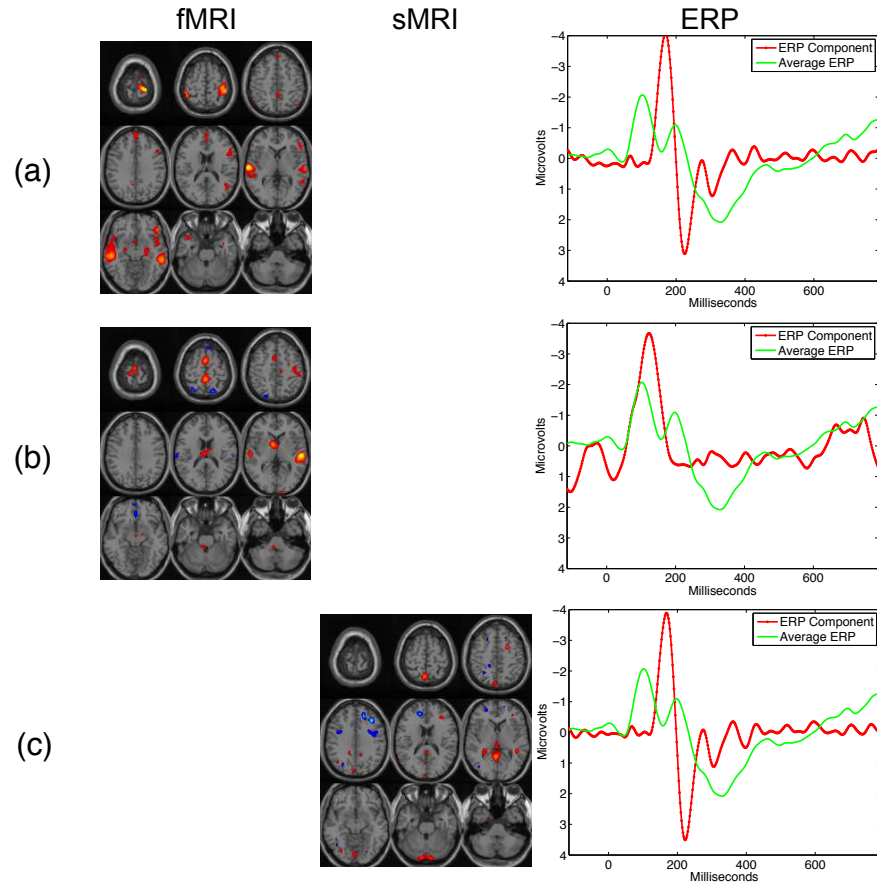


FIG. 4.4. Significant and biologically meaningful components generated using jICA using the EBM algorithm for combinations of two modalities. The uncorrected significance values for the components are (a) 0.0031 for the first component and (b) 0.0147 for the second component from the fusion of fMRI and ERP, as well as (c) 0.0044 for the component from the fusion of sMRI and ERP. All spatial maps are Z-maps thresholded at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients.

ness” of the components when bringing a third modality into the analysis, in addition to providing a link across the modalities.

4.4.2 TCCA/TMCCA/TIVA Results and Discussion

Figure 4.5 shows the fusion results using all three modalities using tMCCA-MAXVAR, tMCCA-GENVAR, and tIVA-G. The first thing to note is that since, unlike for jICA, each dataset has its own set of subject covariations, it is possible for a component to not be sig-

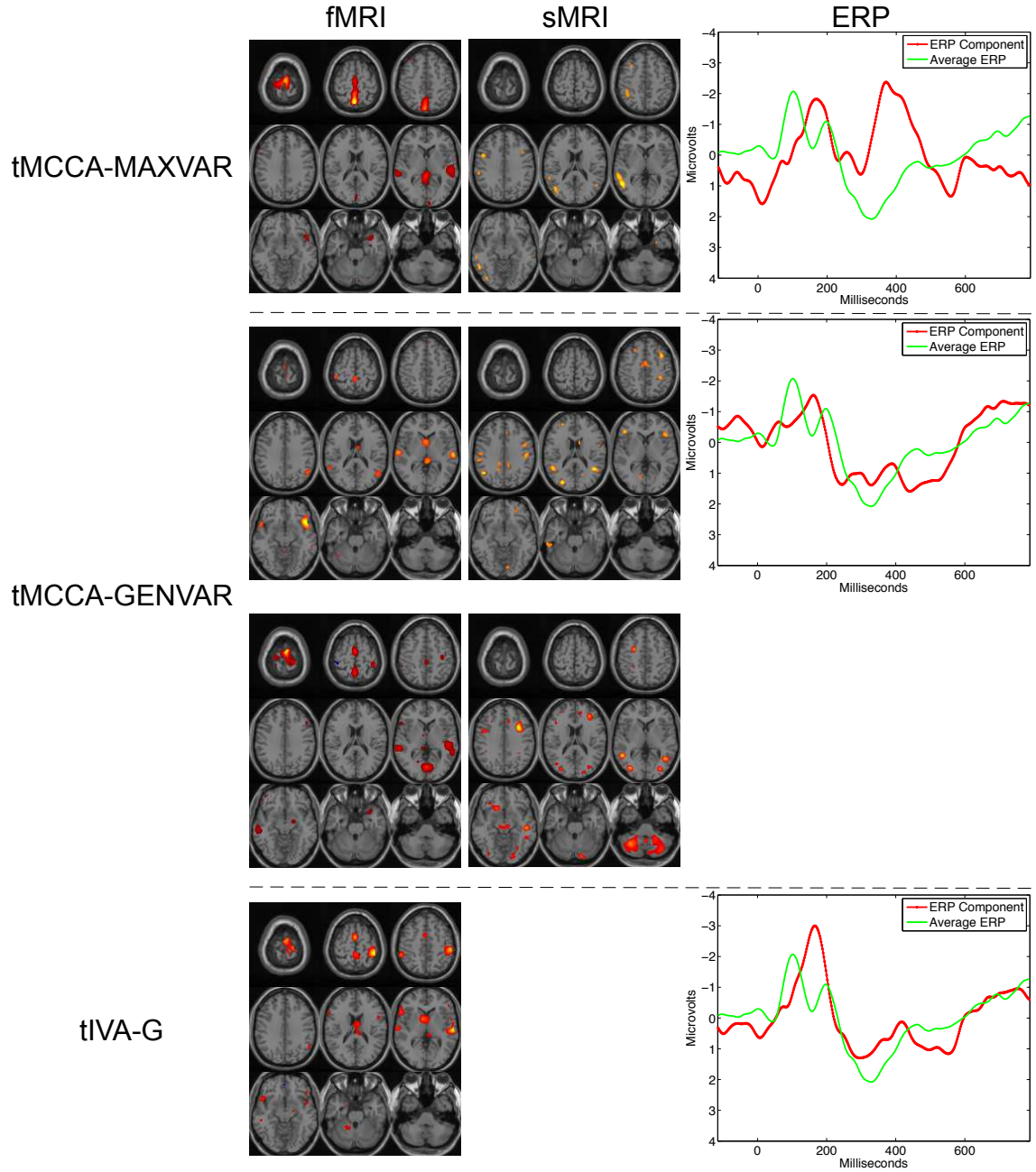


FIG. 4.5. Significant and biologically meaningful components generated from the fusion of all three datasets using tMCCA using the MAXVAR and GENVAR cost functions as well as tIVA-G. The uncorrected significance values for the components estimated using MAXVAR are 0.0140. The uncorrected significance values for the components estimated using GENVAR 0.0134 for the top component and 0.0187 for the second one. For tIVA-G, we have a single component that shows significant group difference with 0.0265. All spatial maps are Z-maps thresholded at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients.

nificant in all three modalities at the same time, as can be seen in Figure 4.5 for tMCCA-GENVAR and tIVA-G. We should also note that when reporting the significance for these fusion results, we report the p -value corresponding to the mean of the significant t -statistics for each component. We see some relations between the component found using jICA, depicted in Figure 4.3(a) and the tIVA-G result, particularly in the ERP, which both seem to characterize the P2/N2 transition. This transition is also characterized in the ERP components found by tMCAA-MAVAR and tMCCA-GENVAR. We also see some similarities between the fMRI portion of these components, particularly the increased sensorimotor and temporal activation of the controls versus the patients. However, in the case of tMCCA-GENVAR, these regions in the fMRI are split between two components. Since both tMCCA-GENVAR and tIVA-G make use of the same cost function with the difference being that tMCCA-GENVAR assumes that the demixing matrices are orthogonal, it is reasonable that the two results are similar. In fact, combining the two significant fMRI components from tMCCA-GENVAR would result in an fMRI component that is nearly identical to the one found using tIVA-G. In addition, since IVA-G is the least constrained approach among the three that we consider here, the fact that the sMRI component is not significant is meaningful, due to the fact that it supports the conclusions reached with the jICA results, namely that the sMRI has the weakest connection to the other modalities.

Figure 4.6 displays the results of the pairwise analysis of the fMRI and ERP data using both tCCA and tIVA-G. We can see that these results are quite similar to each other, highlighting the connection between tCCA and tIVA-G. Again, these results are similar to fusion results when analyzing all three modalities for tIVA-G. However, the significance value is lower for the results using all three modalities than it is when only fusing the fMRI and ERP. These results again imply that the fMRI and ERP data are driving the fusion results with three modalities, with the sMRI providing a more minor role. This, in

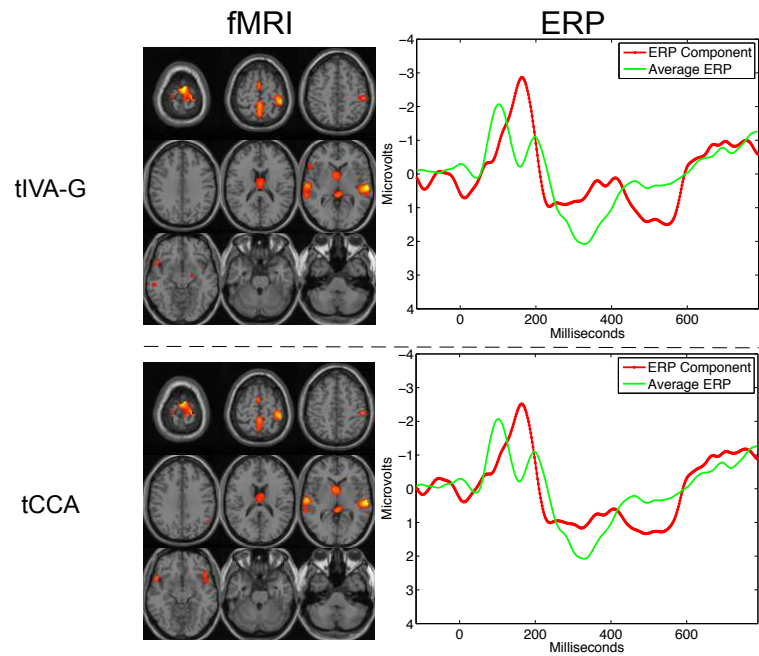


FIG. 4.6. Significant and biologically meaningful components generated using tIVA-G and tCCA with two modalities. The uncorrected significance for components are 0.0130 for the component estimated using tIVA-G and 0.0130 for the component estimated using tCCA. All spatial maps are Z-maps thresholded at $Z=3.5$. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients.

addition to the lower significance values when combining all three modalities, highlights the fact that, though it is attractive to fuse information from all available modalities, there may be penalties—in this case an overall decrease in the sensitivity of the detected group differences—when combining modalities that are less discriminative themselves, see *e.g.*, [114].

4.5 Summary

The data fusion techniques that we describe in Chapter 2 have the potential to greatly facilitate our understanding of neural function and how it is disrupted by diseases such as schizophrenia. However, there are a number of practical issues associated with combining different modalities that may not be apparent initially, which in this application were the limited number of subjects and the unequal contribution of the modalities to the final result. For the given datasets, since the number of subjects is only 36, there is a limit on the performance of tCCA/tMCCA/tIVA. On the other hand, having a harder constraint, such as the one imposed by jICA, may lead to results that are more significant and easier to interpret biologically. However, motivated by the simulation results, we caution against measuring performance using only the statistical significance of the results, since jICA may severely overestimate the t -statistics, while tCCA/tMCCA/tIVA may more accurately estimate them—particularly if the connection between the modalities is weak. Therefore, we may be more confident in the jICA results for the combination of fMRI and ERP, since their link appears to be strong and the assumption that the modalities share a common mixing matrix can average out the effects of additive noise. On the other hand, for the combinations of fMRI and sMRI, sMRI and ERP, and, possibly, the combination of all three, the results found using tCCA/tMCCA/tIVA may be more reasonable.

Chapter 5

QUANTIFYING THE INTERACTION AND CONTRIBUTION OF MULTIPLE DATASETS IN FUSION

The analysis of multiple datasets drawn from the same subjects is possible by either analyzing the datasets jointly as in data fusion, or separately and then combining, as in data integration. However, selecting the optimal method to combine and analyze multiset data is an ever-present challenge. The primary reason for this is the difficulty in determining the optimal contribution of each dataset to an analysis as well as the amount of potentially exploitable complementary information among datasets. In this chapter, we introduce a novel classification rate-based technique to unambiguously quantify the contribution of each dataset to a fusion result as well as to facilitate direct comparisons of fusion methods on real data and apply IVA to multiset fusion. We apply this classification rate-based technique to fMRI data from 121 patients with schizophrenia and 150 healthy controls during the performance of three tasks. Through this application, we find that though we achieve optimal performance by exploiting all tasks, each task does not contribute equally to the result and this framework enables effective quantification of the value added by each task. Our results also demonstrate that data fusion methods are more powerful than data integration methods, with the former achieving a classification rate of 73.5 percent and the latter achieving one of 70.9 percent, a difference which we show is significant, when we analyze

all three tasks together. Finally, we also show that IVA, due to its flexibility, has equivalent or superior performance compared with the popular data fusion method jICA.

5.1 Introduction

The collection of data from multiple sensors raises the issue of how best to utilize each of these datasets in order to maximize the use of available information for the given task. However, there are several issues that prevent straightforward selection of the ideal method when analyzing multiple real world datasets, these include: quantification of the additive value of each dataset, determining the combination of datasets that achieves the best performance, and quantification of the joint information between datasets within an analysis.

We propose a novel classification rate-based technique to assess the performance of different multiset analysis methods for various combinations of datasets. We should note that perfect classification is not the goal of this work, especially since the patients with schizophrenia were receiving antipsychotic and/or mood stabilizing medication during the scanning sessions, thus making it difficult to attribute observed differences to the disease or to the medication, see *e.g.*, [113, 115, 116]. Instead, our goal is to enable unambiguous quantification of the additive value of each dataset, determination of the combination of datasets that achieves the greatest performance, and quantification of the interaction among datasets within an analysis. Note that the proposed technique differs fundamentally from the nonparametric prediction, activation, influence, and reproducibility resampling (NPAIRS) framework [117], whose goal is to assess how well a model generalizes to new data rather than how useful a dataset is to a fusion analysis. We apply this technique to fMRI data from three tasks: AOD, SIRP, and SM, drawn from 121 patients with schizophrenia and 150 healthy controls, described in Chapter 2.

5.2 Classification Procedure

Since there is no ground truth for the components or subject covariations, we propose to use classification rate to determine the relative performance of different techniques on real multitask fMRI datasets. We use this as our metric for two reasons: first, classification rate has an unambiguous meaning that can be defined for different combinations of datasets and analysis techniques, and second, classification rate effectively encapsulates the total discriminative power between patients and controls. A diagram of the classification procedure is shown in Figure 5.1 and the process is described below.

We will describe the classification process for a single dataset using ICA, since the extension to multiple datasets and different analysis techniques is, as we will show, straightforward. The first step is randomly resampling 190 subjects from the original fMRI feature dataset, \mathbf{X} , in order to produce a dataset to train the classifier, $\mathbf{X}_{\text{Train}}$. The feature data from the remaining 81 subjects is formed into \mathbf{X}_{Test} , which will be used to test the trained classifier. In order to reduce any bias introduced in the random selection of subjects, the proportion of patients and controls is kept the same in \mathbf{X} , $\mathbf{X}_{\text{Train}}$, and \mathbf{X}_{Test} . In the next step, PCA, using the order specified using PCA-CCA, then ICA-EBM, is performed on $\mathbf{X}_{\text{Train}}$. Following ICA-EBM, a two-sample t -test is performed on each column of the estimated subject covariations, $\hat{\mathbf{A}}_{\text{Train}}$, and those that are declared significant, *i.e.*, $p < 0.05$, as well as the corresponding spatial maps are formed into $\check{\mathbf{A}}_{\text{Train}}$ and $\check{\mathbf{S}}_{\text{Train}}$, respectively.

The final stage of the classification process begins with the training of a classifier, such as a radial basis function kernel support vector machine (KSVM) using the rows of $\check{\mathbf{A}}_{\text{Train}}$. Next, $\check{\mathbf{A}}_{\text{Test}}$ is found by regressing $\check{\mathbf{S}}_{\text{Train}}$ onto \mathbf{X}_{Test} and is used to test the classifier produced using $\check{\mathbf{A}}_{\text{Train}}$. The value of the classification parameter, such as the kernel parameter for KSVM, is selected by computing the average classification rate for 800 independent Monte-Carlo resamplings of the data with different parameter values and finding the value with the

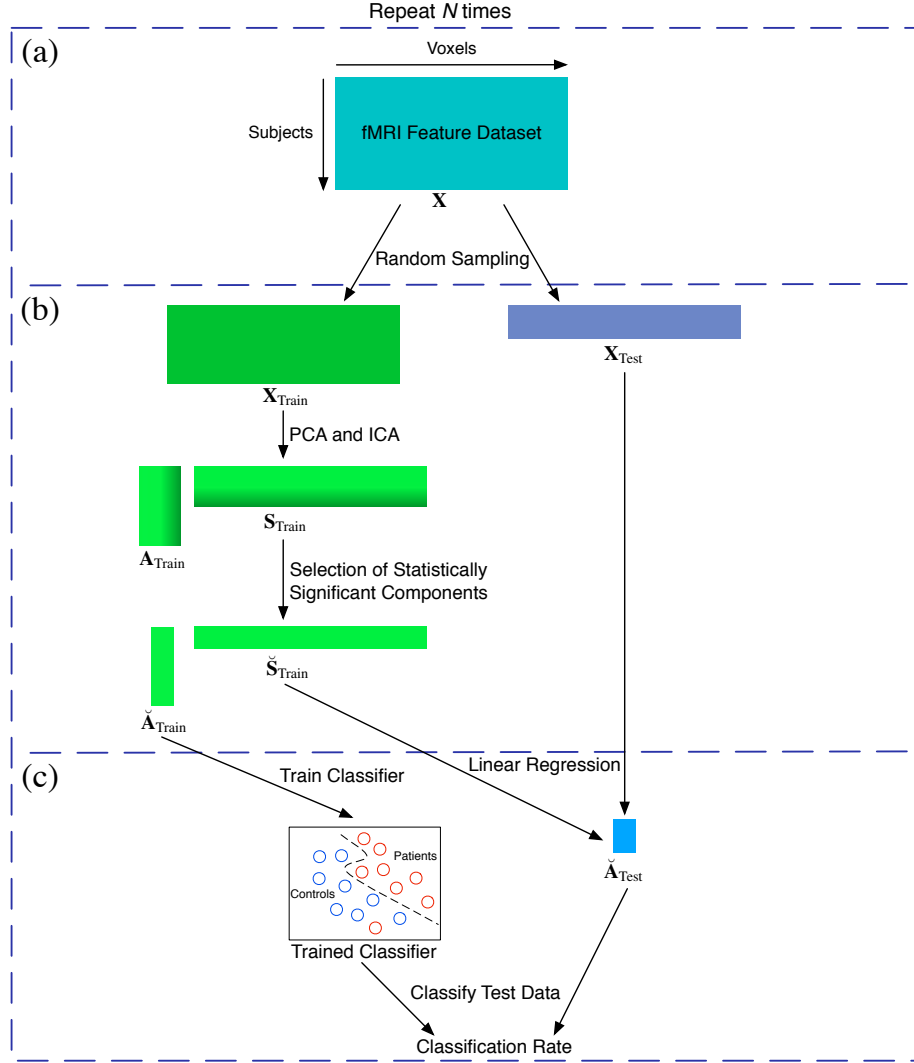


FIG. 5.1. Classification process for a single feature dataset. For the case where multiple datasets are analyzed, the ICA step is replaced with either jICA or IVA, performed on the concatenated feature datasets or the collection of feature datasets, respectively. The procedure is as follows: (a) the data is split into a training set, X_{Train} , and a test set, X_{Test} . (b) the training dataset is dimension reduced using PCA, ICA is run, and the discriminatory components, \hat{S}_{Train} , and corresponding subject covariations, \hat{A}_{Train} , are selected. (c) in the final stage, \hat{A}_{Train} is used to train the classifier, \hat{S}_{Train} is regressed onto X_{Test} producing \hat{A}_{Test} , and \hat{A}_{Test} is used to test the classifier. This process is repeated N times and the mean classification rate is evaluated. Note that for IVA and well as the data integration technique, there are dataset-specific subject covariations $\hat{A}_{\text{Train}}^{[k]}$ and $\hat{A}_{\text{Test}}^{[k]}$; however, for IVA the subject covariations are derived by fusing information across all datasets whereas for the data integration technique the subject covariations are extracted from each dataset individually.

highest average classification rate. Once the optimal value of the classification parameter is selected, the classification procedure is run on N independent Monte-Carlo resamplings of the data and the mean classification rate is computed. In our results, we use 200 as the value of N .

For jICA, the process is identical to the one described above, except \mathbf{X} is composed of the concatenated feature datasets. With IVA, the process is similar to the one described above, except the two-sample t -test is run on every column of all estimated mixing matrices, $\hat{\mathbf{A}}_{\text{Train}}^{[k]}$, and the columns showing a significant difference between the patients and controls are all concatenated together when training the classifier. In order to compare these two data fusion methods with a technique that does not analyze the datasets jointly, *i.e.*, data integration, we took the same setup from IVA, but replaced the IVA step with a separate ICA analysis for each feature dataset individually. In this work, we use the IVA using a multivariate Gaussian and Laplacian (IVA-GL) algorithm [118], since it is an effective IVA algorithm for analysis of fMRI data, see *e.g.*, [118, 119]. Note that to facilitate comparisons between methods and remove possible confounding from the use of different orders for different methods, we use the same order, 24, estimated using PCA-CCA, in the PCA step for all methods. Possible effects of the order on classification performance were studied by repeating the ICA analyses for different orders and, as we show in Section 5.2.2, while order-selection is important for most applications, we observe no statistically significant differences, after a Bonferroni correction, between the classification rates at different orders.

Note that if the classifier requires the determination of a parameter, such as the value of kernel parameter for KSVM, the selection of the optimal value is done for each method individually. One final note regarding the computational complexity, since the number of samples is much larger than the PCA order and the number of datasets, *i.e.*, $V \gg C, K$,

by far the most time consuming step in the proposed framework is the jICA/IVA step. The computational complexity is approximately $O(KVC^2)$ per iteration and approximately $O(KVC^2 + K^2VC)$ per iteration for jICA using EBM and IVA-GL, respectively.

5.2.1 Number of Monte-Carlo Resamplings

In order to evaluate the validity of using 200 Monte-Carlo resamplings, we evaluate the classification rate using KSVM and IVA on the combination of AOD and SIRP datasets for different numbers of Monte-Carlo resamplings and display the results in Figure 5.2.

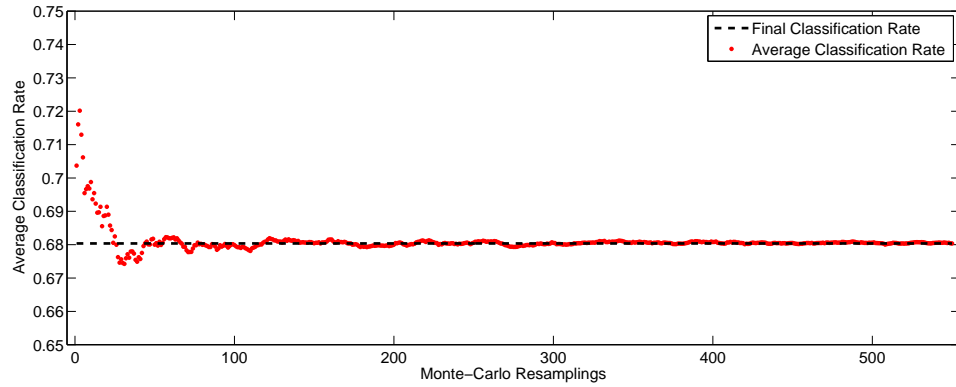


FIG. 5.2. Average classification rate using KSVM and IVA on the combination of AOD and SIRP datasets as a function of Monte-Carlo resamplings.

From Figure 5.2, it is clear that the classification rate converges to a stable result fairly rapidly as a function of the number Monte-Carlo resamplings and the final classification rate is nearly identical for any number of resamplings greater than approximately 150. We should note that though we only show the result of one combination of classification procedure, analysis method, and combination of datasets, we see very similar trends in the other combinations as well.

5.2.2 Order-Selection

In order to probe the effect of the order on the classification results, we rerun the proposed framework 1,000 times, using 800 runs for training and testing and 200 runs for validation, on the AOD dataset using ICA-EBM for different values of C . We study this scenario due to the fact that we want to investigate purely the effects of selecting a different PCA order, thus we want to keep the setup as simple as possible. We use the KSVM as the classifier, since it scales well with the dimension of training space, and train the KSVM parameter separately for each value of C . We display the results in the Figure 5.3.

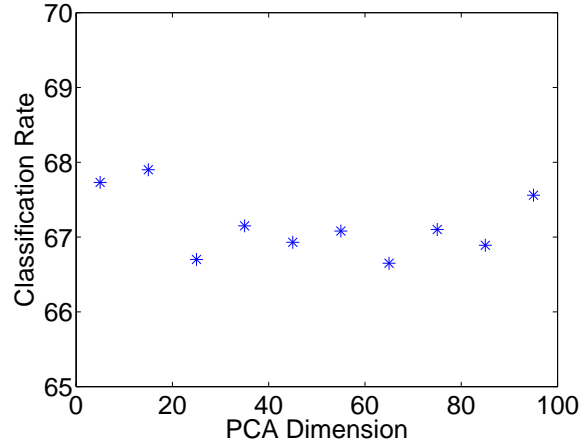


FIG. 5.3. Average classification results using ICA-EBM and KSVM on the AOD dataset for different orders.

We expect the number of significant components and therefore the dimensionality of the space in which we classify to generally increase with C , thus making the classification problem easier. However, we do not see this in general in Figure 5.3, which may possibly indicate overfitting for higher values of C . Though we show only one result here, the trends are fairly consistent with those that we observe for the ICA analyses of the other datasets, *i.e.*, no consistent general trends as a function of C . We see some variation in the classification rate as a function of C , however it is hard to determine any general trends

and the classification rate seems fairly consistent for orders above 15. In order to probe the differences in classification rate that we observe in Figure 5.3, we run a two-sample t -test comparing the highest classification rate achieved, *i.e.*, using $C = 15$, with all other values of the classification rates achieved for different values of C and display the results in Table 5.1.

C	5	15	25	35	45	55	65	75	85	95
p -value	NS	–	0.011	NS	NS	NS	0.009	NS	0.039	NS

Table 5.1. Significance of the difference in classification rate using KSVM of ICA-EBM using a PCA order, C , of 15 compared to other values of C . We assess statistical significance through a two-sample t -test performed on the classification rates obtained for 200 independent resamplings of the fMRI feature datasets. If the difference corresponds to $p \geq 0.05$, the result is marked “not significant” (NS). Note that there is no correction for multiple comparisons and none of the observed differences would pass a Bonferroni correction (0.0056 for a p -value of 0.05).

From Table 5.1, we can see that though we see some significance in terms of the differences that we observe in Figure 5.3, they are mostly minor and would not survive a correction for multiple comparisons. These results do motivate a future analysis where we tune the value of C for each method individually. Such an investigation could potentially be interesting, however, this would raise the issue of whether the different value of C or the fusion method itself was the source of the differences in classification rate, thus distracting from the fundamental goal of this chapter, *i.e.*, quantifying the contribution of each dataset to a fusion result as well as facilitating direct comparisons of fusion methods.

5.3 Results

We display the classification results for the individual multitask datasets as well as their combinations, using KSVM in Figure 5.4. Several noteworthy trends can be observed in Figure 5.4. First, we can see that equivalent or improved classification rate is achieved when multiple datasets are jointly analyzed than when only a single dataset is analyzed.

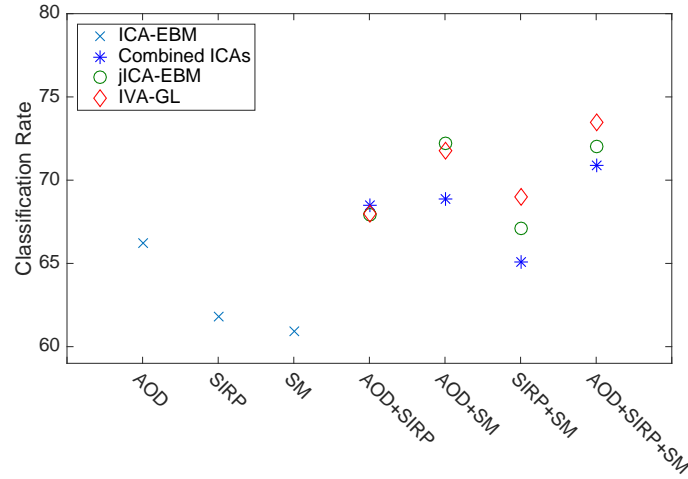


FIG. 5.4. Average classification results using KSVM for individual datasets and combinations of datasets using either data fusion, with jICA and IVA-GL, or data integration using combined ICAs. The first three columns from the left refer to the case where only one dataset is analyzed. The fourth, fifth, and sixth columns from the left refer to combinations of two datasets. The rightmost point shows the classification performance when all three task datasets are jointly analyzed. Note that error bars are omitted for clarity, since the largest value of the standard error is 0.0035.

This is due to the fact that there is more total discriminative power available in a combination of multiple datasets than there is in any of the datasets individually. Additionally, the AOD dataset has the highest classification rate of all of the individual datasets, while the SM dataset has the lowest. This means that, when analyzed separately, the AOD dataset is best able to differentiate between patients and controls, while the SM dataset has more difficulty. We also note that there is a clear advantage to be found through the use of data fusion, *i.e.*, jICA and IVA-GL, over the use of data integration, *i.e.*, combined ICAs, in all cases, except where only the AOD and SIRP datasets are used, where equivalent performance is achieved. This shows that the results for jICA and IVA-GL are more than the “sum of their parts,” hence, emphasizing importance of allowing datasets to fully interact with each other in an analysis [4, 112]. Additionally, in all cases IVA-GL performs better than or is statistically equivalent to jICA, showing the advantages of using a less constrained model when the signal-to-noise ratio is relatively high [13]. The strong performance of

Datasets	Significance of Difference (p -values)	
	IVA-GL to Combined ICAs	IVA-GL to JICA
AOD and SIRP	NS	NS
AOD and SM	1.0×10^{-10}	NS
SIRP and SM	1.4×10^{-16}	9.5×10^{-5}
All Three Datasets	2.3×10^{-9}	8.5×10^{-4}

Table 5.2. Significance of the difference in classification rate using KSVM of IVA-GL compared to the data integration technique, combined ICAs, and the popular data fusion method, jICA. Statistical significance is assessed through a two-sample t -test performed on the classification rates obtained for 200 independent resamplings of the fMRI feature datasets. If the difference corresponds to $p \geq 0.05$, the result is marked “not significant” (NS). Note that there is no correction for multiple comparisons, but all significant differences would remain so after the conservative Bonferroni correction.

IVA-GL when compared with the ICA-based techniques is noteworthy since the algorithm used for the ICA-based techniques was EBM, which is able to fit a much larger range of latent source distributions than IVA-GL, thus showing the power of methods that exploit complementary information across datasets. Since IVA-GL is best able to summarize the total discriminative power of a combination of datasets, we can use the difference in classification rate between a two dataset analysis and the three dataset analysis to assess the additional contribution of the third dataset, beyond what was provided by the other two, to the fusion result. For example, the additional contribution of the SM dataset is the difference between the classification rate of IVA-GL using the AOD and SIRP datasets and the classification rate of IVA-GL using all three datasets. Based on this, the additional contribution of AOD is 4.5 percent, the additional contribution of SIRP is 1.7 percent, and the additional contribution of SM is 5.5 percent. Thus, the SM and AOD datasets contribute much more exploitable information than the SIRP dataset. In order to provide a qualitative assessment of the differences observed in Figure 5.4, 2-sample t -tests were performed on the classification rates and the results are shown in Table 5.2.

As can be seen from Table 5.2, all of the major differences observed visually in Figure 5.4 are statistically significant. No correction for multiple comparisons is performed,

however we note that all of the observed differences remain statistically significant even after the conservative Bonferroni correction. An interesting result that can be seen in both Figure 5.4 and Table 5.2 is that for the combination of AOD and SIRP feature datasets, there is no statistical difference in the classification rates of IVA-GL compared with that of the combined ICAs. This indicates that there is no measurable difference in performing data fusion or data integration for this combination of datasets. These results suggest that there is little joint information that is exploited between the AOD and SIRP datasets. On the other hand, there is significant improvement in performance when fusing the AOD and SM datasets, implying that there is a significant amount of joint information between the AOD and SM datasets.

5.3.1 Stability of Conclusions Using Other Classification Techniques

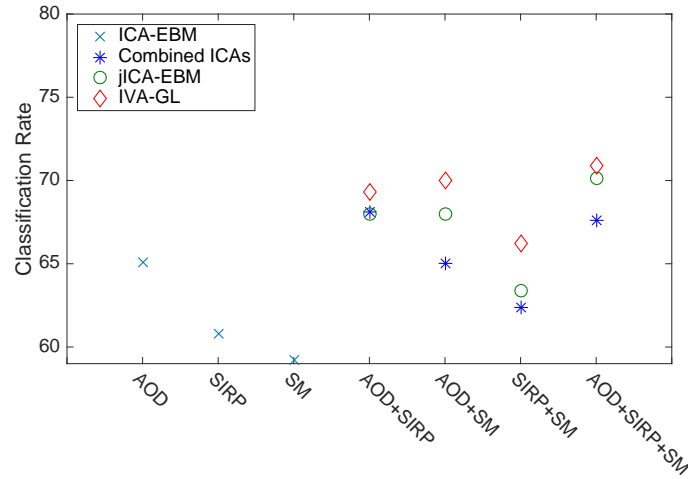


FIG. 5.5. Average classification results using KNN for individual datasets and combinations of datasets using either data fusion, with jICA and IVA-GL, or data integration using combined ICAs. The first three columns from the left refer to the case where only one dataset is analyzed. The fourth, fifth, and sixth columns from the left refer to combinations of two datasets. The rightmost point shows the classification performance when all three task datasets are jointly analyzed. Note that error bars are omitted for clarity, since the largest value of the standard error is 0.0034.

The motivation for using KSVM is its robust performance in cases, similar to ours,

where the number of training samples is limited. To confirm the results that we state in the previous section based on the KSVM, we also repeat our analysis with two other popular fusion methods: K -nearest neighbors (KNN) and Fisher discriminant analysis (FDA). For each method, we optimize the classification parameters, *i.e.*, σ for the KSVM and K for KNN, with the same training/test data consisting of 800 independent runs of each of the methods on each of the combinations of resampled datasets. We use the same validation set, consisting of 200 additional runs on separate resampled datasets, for each method and report the average classification rate as well as the results of two-sampled t -tests performed on the classification rates of the different methods below.

Datasets	Significance of Difference (p -values)	
	IVA-GL to Combined ICAs	IVA-GL to JICA
AOD and SIRP	0.0062	0.0038
AOD and SM	5.5×10^{-25}	1.2×10^{-5}
SIRP and SM	1.5×10^{-19}	1.3×10^{-11}
All Three Datasets	8.3×10^{-13}	NS

Table 5.3. Significance of the difference in classification rate using KNN of IVA-GL compared with the data integration technique, combined ICAs, and the popular data fusion method, jICA. Statistical significance is assessed through a two-sample t -test performed on the classification rates obtained for 200 independent resamplings of the fMRI feature datasets. If the difference corresponds to $p \geq 0.05$, the result is marked “not significant” (NS). Note that there is no correction for multiple comparisons, but all significant differences would remain so after the conservative Bonferroni correction.

From Figures 5.5 and 5.6 as well as Tables 5.3 and 5.4, we note statistically equivalent or improved classification rates when multiple datasets are jointly analyzed than when only a single dataset is analyzed. Additionally, the AOD dataset has the highest classification rate of all of the individual dataset analyses, while the SM dataset has the lowest. We see that, as we note when using KSVM, there is a significant improvement in classification rate, using both KNN and FDA, of IVA-GL compared to the combined ICAs for the combination of AOD and SM. Though we note that there is a statistically significant difference in classification rate, using KNN, of IVA-GL compared to the combined ICAs for the com-

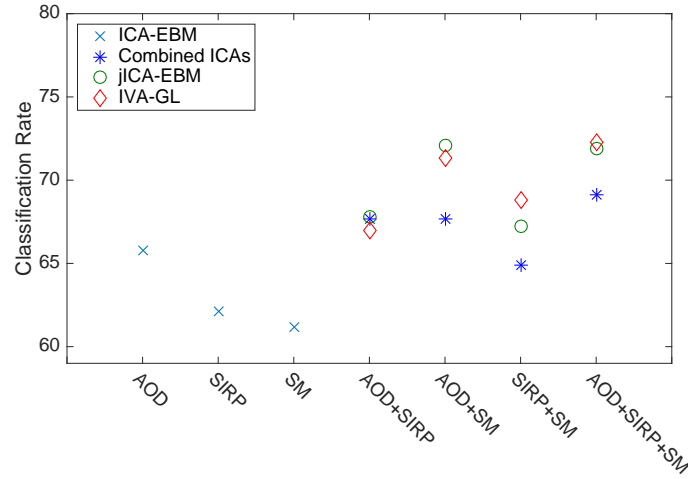


FIG. 5.6. Average classification results using FDA for individual datasets and combinations of datasets using either data fusion, with jICA and IVA-GL, or data integration using combined ICAs. The first three columns from the left refer to the case where only one dataset is analyzed. The fourth, fifth, and sixth columns from the left refer to combinations of two datasets. The rightmost point shows the classification performance when all three task datasets are jointly analyzed. Note that error bars are omitted for clarity, since the largest value of the standard error is 0.0037.

combination of AOD and SIRP, the difference is much less significant than for the combination of AOD and SM. We also note that there is an advantage to be found through the use of data fusion, *i.e.*, jICA and IVA-GL, over the use of data integration in most cases or statistically equivalent performance is achieved. A final and important trend that we see is that in all cases IVA-GL performs better than or equivalent to jICA.

5.3.2 Exploring the Effects of Fusion

In order to probe the differences between combining the AOD and SIRP datasets versus combining the AOD and SM datasets, we run both IVA-GL and the data integration technique on all of the data 10 times and select a run from each method for comparison using the MST-based method described in Section 2.4. The reason that we use the components from this run is that each of the results of a method on the resampled data may be seen as rough approximations of the results of the same method on the whole dataset, since we

Datasets	Significance of Difference (p -values)	
	IVA-GL to Combined ICAs	IVA-GL to JICA
AOD and SIRP	NS	NS
AOD and SM	6.4×10^{-13}	NS
SIRP and SM	1.8×10^{-16}	8.3×10^{-4}
All Three Datasets	8.0×10^{-12}	NS

Table 5.4. Significance of the difference in classification rate using FDA of IVA-GL compared with the data integration technique, combined ICAs, and the popular data fusion method, jICA. Statistical significance is assessed through a two-sample t -test performed on the classification rates obtained for 200 independent resamplings of the fMRI feature datasets. If the difference corresponds to $p \geq 0.05$, the result is marked “not significant” (NS). Note that there is no correction for multiple comparisons, but all significant differences would remain so after the conservative Bonferroni correction.

greatly reduce the dimension of the datasets using PCA prior to performing the analysis. Using the run that the MST-based method selects, we find the p -values by performing a two-sample t -test on each column of the estimated mixing matrices of that run, where one group is the subject covariations of the controls and the other is the subject covariations of the patients. Those p -values that pass a significance threshold of $p < 0.05$ are declared statistically significant and we align the corresponding Z-scored components between the data fusion and data integration techniques. Two components are considered aligned if they have a sample Pearson correlation coefficient above 0.5, which we find matches with the results produced through visual alignment. We show the results in Figures 5.7 and 5.8.¹

There are two trends that become clear from the comparison of Figures 5.7 and 5.8. The first is that the proportion of components that are declared significant in the IVA-GL results, which are not seen in the ICA results, is much higher for the combination of AOD and SM than it is for AOD and SIRP. This would imply that the IVA-GL decomposition shows greater difference for the combination of AOD and SM, *i.e.*, there is greater joint information that is exploited. This helps explain the significant improvement in the

¹An exploration of how the IVA and combined ICAs relate to the jICA results can be found in Figures 5.9 and 5.10.

classification rate of the data fusion technique over the data integration technique for the combination of AOD and SM as well as the reason for its absence in the combination of AOD and SIRP. We should note that the regions found to be significantly different in patients and controls that do not correspond to the ventricles and are found for both IVA and the individual ICAs in Figure 5.7 correspond to: medial visual cortex, fronto-insular cortex, thalamus, cerebellum, left fronto-parietal cortex, motor cortex, visual cortex, and sensorimotor cortex. These regions are known to differentiate patients with schizophrenia from controls, see *e.g.*, [25, 27, 120–123]. Similarly, the regions that activate significantly differently in patients and controls in Figure 5.8 correspond to the: medial visual cortex, DMN, motor cortex, thalamus, left fronto-parietal cortex, cerebellum, visual cortex, and sensorimotor cortex. These regions are also known to be impacted by schizophrenia, see *e.g.*, [25, 27, 120–124], consistent with the idea of cortico-cerebellar-thalamic-cortical circuit disruption or cognitive dysmetria in schizophrenia [125], thus increasing our overall confidence in the results of this analysis. In addition to these, we also have differentiating components with spatial activations corresponding to the posterior parietal lobe in Figure 5.7 and the temporal lobe and right fronto-parietal lobe in Figure 5.8 in the IVA results and not in the ICA results. These regions are known to differentiate patients from controls, see *e.g.*, [25, 27], and they, coupled with the higher, on average, significance values for the components, show the advantages of a fusion analysis over individual analyses.

The second trend that we see from both Figure 5.7 and Figure 5.8 is that there are multiple significant components that we find in the single dataset analyses but not in the fusion results. This may be due to the fact that the fusion results report joint components, thus they may mask weaker individual components [17]. The regions that we find in the ICA results and not in the IVA results correspond to the cerebellum, DMN, fronto-parietal cortex, and medial-visual cortex in Figure 5.7 and to the cerebellum, DMN, temporal cor-

tex, and frontal lobe in Figure 5.8. These regions are known to be affected in patients with schizophrenia, see *e.g.*, [25, 27, 120, 124], and their absence from the IVA results demonstrates the importance of performing a comparison between the fusion results with the ICA results in order to determine the joint as well as dataset-specific information.

These results also motivate the use of the proposed framework for automated feature selection. Determining optimal features from different modalities/tasks is a continuing challenge in many data fusion studies, see *e.g.*, [6, 31, 34, 55, 126, 127]. If the goal of a fusion study is the determination of differences between a set of groups, then the technique developed in this chapter can be used for automated selection of optimal features based on the combination of features that achieves the highest classification rate.

5.4 Summary

A fundamental issue associated with analyzing data from multiple sources is how to most effectively combine datasets from different subjects and tasks. In this chapter, we describe a data-driven classification procedure to determine the performance of different multitask analysis techniques as well as the discriminatory information that a new dataset introduces beyond what already exists in other datasets. We apply this new method to real multitask fMRI data drawn from patients with schizophrenia as well as healthy controls. Through this application, we find that though we achieve the best performance by analyzing all of tasks, each task does not contribute equally to the result and we quantify the value added by each task. Our results show that data fusion methods, through their exploitation of shared information across datasets, are more powerful than data integration methods, which do not take advantage of such information. We also show that the more flexible IVA-GL has equivalent or superior performance compared with the popular data fusion method jICA. Finally, it is important to note that the proposed framework can be applied not only to fMRI

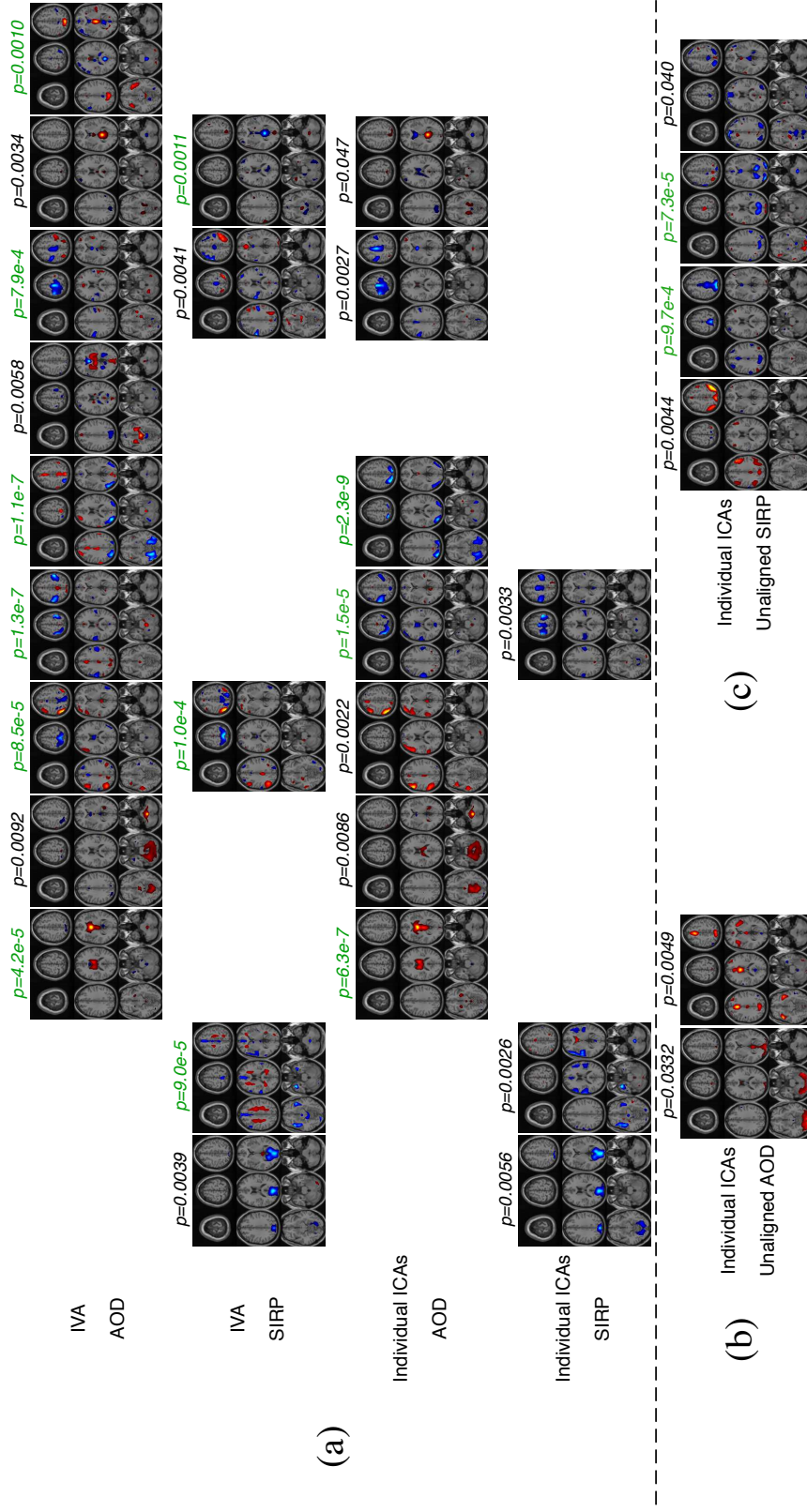


FIG. 5.7. Statistically significant components for the combination of the AOD and SIRP datasets. (a) the significant components obtained through the use of data fusion, using IVA-GL, are shown in the first two rows. The third and fourth rows contain those significant components obtained through the use of data integration, using ICA-EBM, that have a correlation above 0.5 with the components obtained using IVA-GL. The aligned components are in the same column. Those components obtained using ICA-EBM that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (b) and (c) for the AOD and SIRP datasets, respectively. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients. The p -values for each component are located above the corresponding spatial map and those that remain significant after a Bonferroni correction are displayed in green. All spatial maps are Z-maps thresholded at $Z=2.7$.

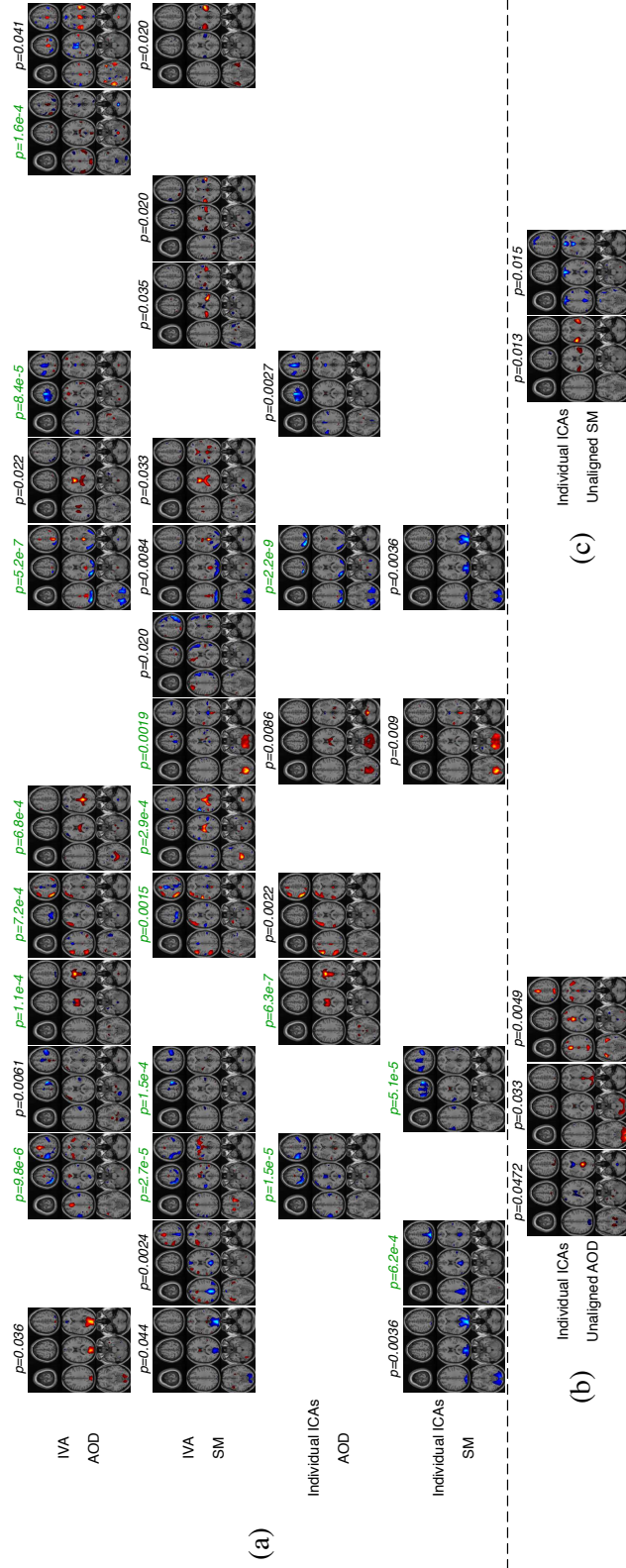


FIG. 5.8. Statistically significant components for the combination of the AOD and SM datasets. (a) the significant components obtained through the use of data fusion, using IVA-GL, are shown in the first two rows. The third and fourth rows contain those significant components obtained through the use of data integration, using ICA-EBM, that have a correlation above 0.5 with the components obtained using IVA-GL. The aligned components are in the same column. Those components obtained using ICA-EBM that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (b) and (c) for the AOD and SM datasets, respectively. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients. The p -values for each component are located above the corresponding spatial map and those that remain significant after a Bonferroni correction are displayed in green. All spatial maps are Z-maps thresholded at $Z=2.7$.

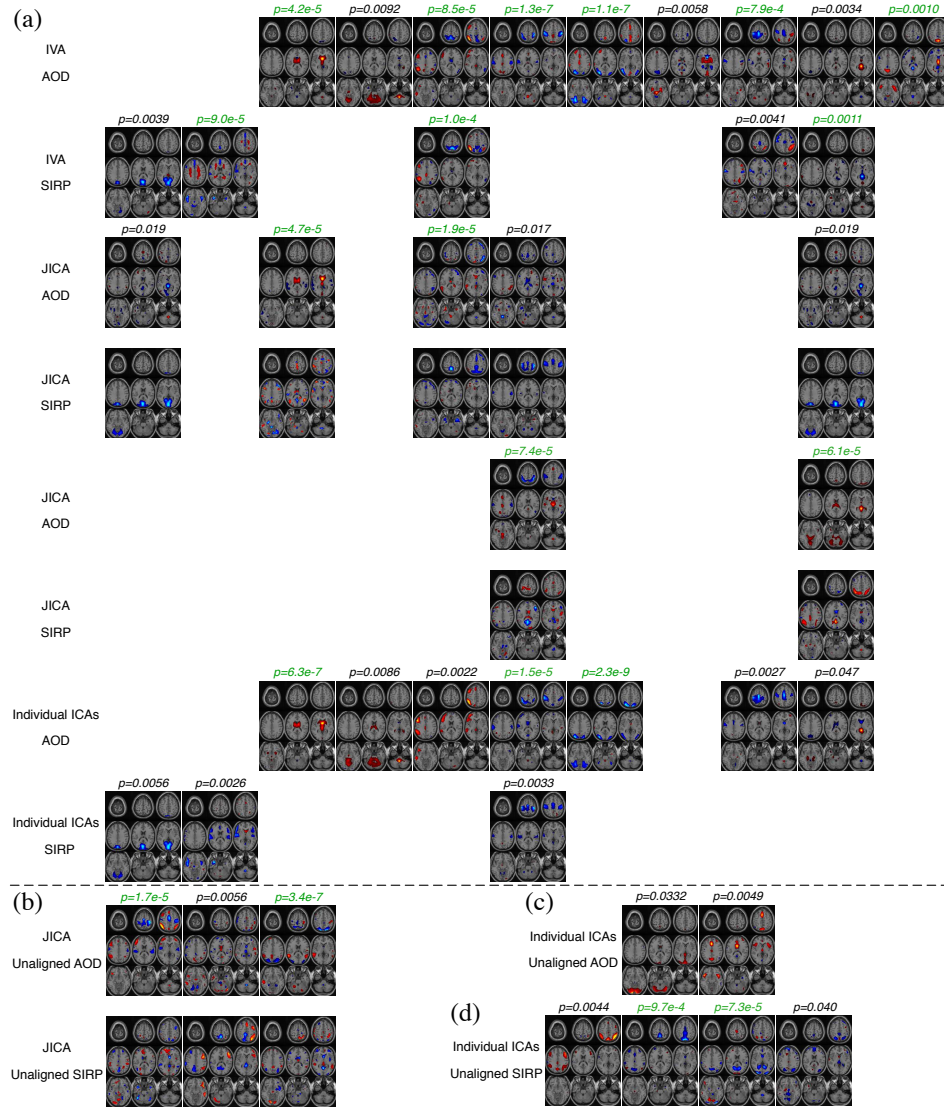


FIG. 5.9. Statistically significant components for the combination of the AOD and SIRP datasets. (a) the significant components obtained through the use of data fusion, using IVA-GL, are shown in the first two rows. The third through sixth rows contain those significant components obtained through the use of jICA that have a correlation above 0.5 with the components obtained using IVA-GL. The aligned components are in the same column. Note that due to the assumption by jICA of a common mixing matrix for all datasets, multiple spatial maps are considered a single component with one corresponding p -value. The seventh and eight rows contain the significant components obtained through the use of data integration, using ICA-EBM, that have a correlation above 0.5 with the components obtained using IVA-GL. The components obtained using jICA that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (b). Additionally, those components obtained using ICA-EBM that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (c) and (d) for the AOD and SIRP datasets, respectively. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients. The p -values for each component are located above the corresponding spatial map and those that remain significant after a Bonferroni correction are displayed in green. All spatial maps are Z-maps thresholded at $Z=2.7$.

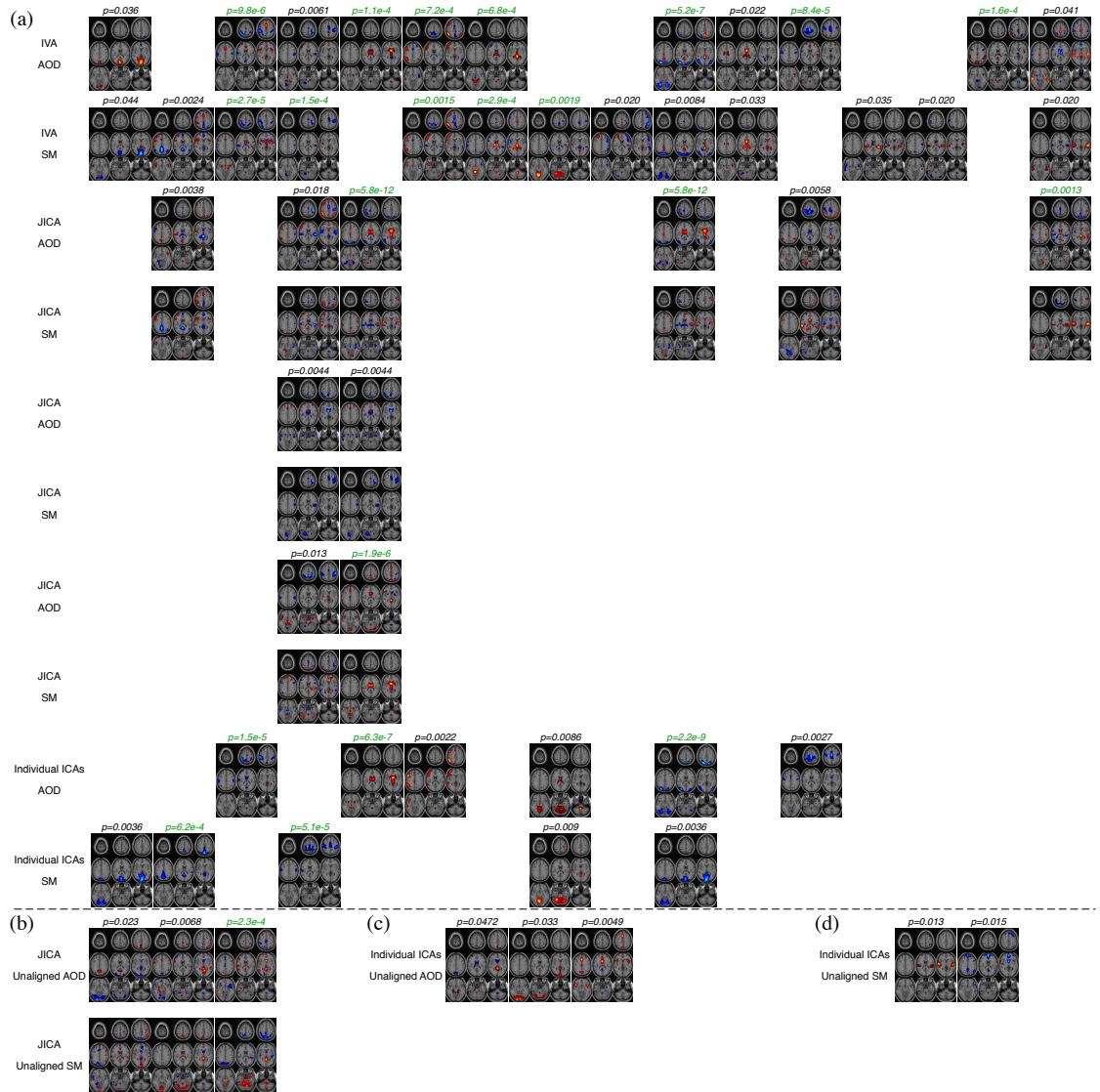


FIG. 5.10. Statistically significant components for the combination of the AOD and SM datasets. (a) the significant components obtained through the use of data fusion, using IVA-GL, are shown in the first two rows. The third through sixth rows contain those significant components obtained through the use of jICA that have a correlation above 0.5 with the components obtained using IVA-GL. The aligned components are in the same column. Note that due to the assumption by jICA of a common mixing matrix for all datasets, multiple spatial maps are considered a single component with one corresponding p -value. The seventh and eight rows contain the significant components obtained through the use of data integration, using ICA-EBM, that have a correlation above 0.5 with the components obtained using IVA-GL. The components obtained using jICA that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (b). Additionally, those components obtained using ICA-EBM that do not have a correlation above 0.5 with any of the components obtained using IVA-GL are shown in the (c) and (d) for the AOD and SM datasets, respectively. Note that we flip the maps such that the activation (red and orange) represents an increase in controls over patients and deactivation (blue) corresponds to a decrease in controls versus patients. The p -values for each component are located above the corresponding spatial map and those that remain significant after a Bonferroni correction are displayed in green. All spatial maps are Z-maps thresholded at $Z=2.7$.

data from multiple tasks, but also to explore the interactions among multimodal data, such as the data that we analyze in Chapters 3 and 4, thus enabling quantification of the additive value of each modality in the analysis.

Chapter 6

VISUALIZING THE CHANGES IN BIOMARKERS FROM FUSION

In Chapter 5, we introduce a classification technique that can be used to quantify the value of an fMRI dataset to a fusion result. However, the proposed method only quantifies the comparative amount of joint discriminatory information and does not summarize how the decomposition itself changes. In order to address this point and to avoid the component alignment step used in Section 5.3.2, in this chapter we develop a new visualization technique, global difference maps (GDMs), and demonstrate how they can be used to visually highlight the changes from fusion and quantify the discriminatory power of each dataset within a decomposition. We apply this technique to multitask fMRI data drawn from 247 subjects, 109 patients with schizophrenia and 138 healthy controls. Through this application, we find that IVA-GL extracts components that generally represent the common brain regions across the tasks, but can mask regions associated with only a subset of the datasets. We also find that the use of GDMs facilitates comparisons across analysis methods in terms of: the significance of the results, how the results are related to certain behavioral scores, and how the discriminatory brain regions change between methods.

6.1 Motivation for and Generation of a GDM

As seen at the end of Chapter 5, for a given analysis technique, there are multiple statistically significant biomarkers that are extracted, thus making a summarization and comparison of different techniques difficult. For this reason, we propose to summarize the performance of a given method through the use of GDMs. We should note that GDMs can be used to augment the strategy developed in Chapter 5 in several ways: first, GDMs provide the discriminatory power of each dataset within a joint decomposition, unlike the former classification technique that summarizes the total discriminatory power of the whole joint decomposition, thus enabling greater insight into the role that each dataset plays. Additionally, GDMs enable visualization of the regions in the brain that have the greatest discriminatory power enabling a visualization of how a decomposition changes by moving from one model to another or including a new dataset in the analysis. We describe the construction of a GDM next.

For the B Z-scored, statistically significant, at $p < 0.05$, components for each dataset, $\hat{\mathbf{s}}_b^{[k]}, 1 \leq m \leq B$, we construct the GDM for that method and dataset, $\hat{\mathbf{s}}_{\text{GDM}}^{[k]}$, as follows

$$\hat{\mathbf{s}}_{\text{GDM}}^{[k]} = \sum_{b=1}^B \frac{t_b}{\sum_{m=1}^B t_m} \hat{\mathbf{s}}_b^{[k]}, \quad k = 1, \dots, K \quad (6.1)$$

where t_b is the t -statistic for the b th biomarker that is positive or made to be positive by multiplying the corresponding subject covariation and biomarker by -1 . Thus, the GDM can be seen as a summary map that describes only the regions that activate significantly differently between patients and controls for a given decomposition and dataset within that decomposition. Each biomarker is scaled by the value of its corresponding t -statistic, so it is weighed more if the component is better able to differentiate between patients and controls. Additionally, we can quantify the discriminative power of a GDM, and thus indirectly the whole decomposition, by generating subject covariations in a nearly identical manner to the

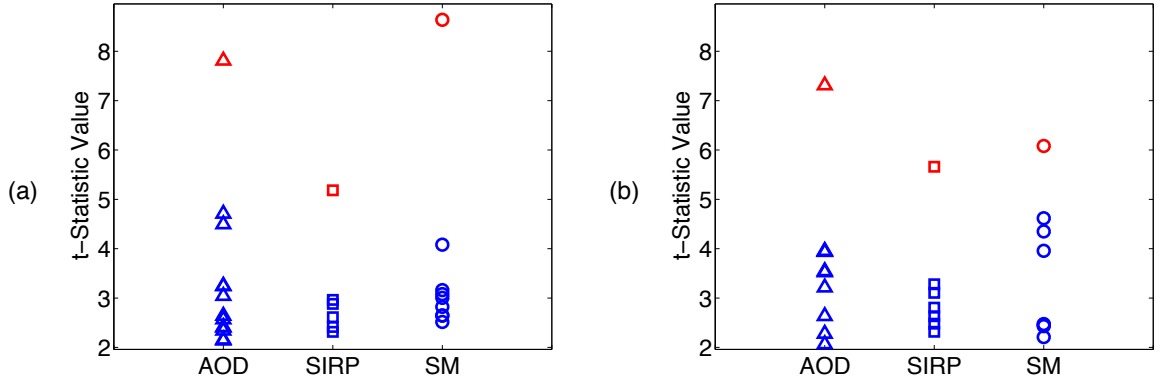


FIG. 6.1. Comparison of the t -statistics from the GDMs compared with the t -statistics of the original biomarkers found using (a) ICA-EBM and (b) IVA-GL for each task. The t -statistics of the GDMs are shown in red, while the original biomarkers are shown in blue.

GDM spatial maps and performing a two-sample t -test on the resulting subject covariations.

The corresponding subject covariations for the GDM are given by

$$\hat{\mathbf{a}}_{\text{GDM}}^{[k]} = \sum_{b=1}^B \frac{t_b}{\sum_{m=1}^B t_m} \hat{\mathbf{a}}_b^{[k]}, \quad k = 1, \dots, K \quad (6.2)$$

where $\hat{\mathbf{a}}_b^{[k]}$ is the subject covariation corresponding to $\hat{\mathbf{s}}_b^{[k]}$. This construction is expected to result in subject covariations that show greater discriminatory power than the original $\hat{\mathbf{a}}_b^{[k]}$, since $\hat{\mathbf{a}}_{\text{GDM}}^{[k]}$ is constructed from only the significant biomarkers. The $\hat{\mathbf{a}}_b^{[k]}$ can be modeled as unit step functions of length equal to the number of controls that are corrupted by Gaussian noise, where higher values of t_b imply lower variances for the noise. Thus, the summation of multiple $\hat{\mathbf{a}}_b^{[k]}$, weighted as in (6.2), should—on average—destructively cancel out some of the noise, thus resulting a lower value for the variance of the noise and a more discriminative subject covariation.

6.1.1 Using GDMs to Summarize Discriminatory Power of a Decomposition

In order to investigate the ability of a GDM to summarize the discriminatory power of a decomposition, we run ICA-EBM and IVA-GL ten times on the multitask fMRI data that

we describe in Chapter 2, drawn from 247 subjects. For each method, we estimate 25 components, corresponding to the maximum order that we find by applying PCA-CCA to the pairwise combinations of the three tasks. We then find the central run for each technique using the MST-based method that we describe in Section 2.4. After selecting the central run, we find the biomarkers and from them calculate GDMs for each task and method. We display the t -statistics of the biomarkers found for each dataset and the corresponding GDMs for ICA-EBM and IVA-GL in Figure 6.1. We also compare the p -value corresponding to the t -statistic from the GDM with the p -value obtained by performing Hotelling’s T -squared test—the multivariate counterpart to the two-sample t -test—with one group being represented by a vector containing the subject covariations of all of the biomarkers from the controls and the other group being represented by a vector containing the corresponding subject covariations from the patients. The results of this comparison are shown in Table 6.1.

	AOD (ICA)	AOD (IVA)	SIRP (ICA)	SIRP (IVA)	SM (ICA)	SM (IVA)
GDM	1.65×10^{-13}	3.74×10^{-12}	4.54×10^{-7}	4.20×10^{-8}	7.52×10^{-16}	4.53×10^{-9}
T -Squared	8.19×10^{-10}	2.48×10^{-9}	3.95×10^{-5}	3.06×10^{-6}	1.15×10^{-14}	1.00×10^{-6}

Table 6.1. Significance in terms of p -values for each combination of dataset and analysis method measured using a 2-sample t -test run on the subject covariations of the GDM as well as Hotelling’s T -squared test. Note that the relative significance of each combination of dataset and analysis method is consistent across the two tests.

From Figure 6.1, we see that the t -statistics from the GDMs are an upper-bound for the t -statistics from their constituent biomarkers. This, combined with the fact that we see the same trend for the significance values in both the 2-sample t -test run on the subject covariations of the GDM as well as Hotelling’s T -squared test, motivate the use of GDMs to summarize the total discriminative power of a dataset within a decomposition. Based on the results shown in Figure 6.1(a) as well as Table 6.1, we can see that, individually, the AOD and SM datasets seem to provide more discriminatory information than the SIRP

dataset, since the significance of the SIRP dataset is much lower than the significance for the AOD and SM datasets. However, we can see from Figure 6.1(b) as well as Table 6.1 that when combined using IVA, the significance of the SIRP dataset increases, while the AOD and SM datasets become less significant. This observation matches the result that we observed in Chapter 4, namely, that there can be a cost, in terms of reduced significance when fusing more discriminatory datasets with less discriminatory datasets.

6.2 Visualizing Fusion

The GDMs extracted from the multitask fMRI data using both ICA-EBM as well as IVA-GL is shown in Figure 6.2. We can see that the discriminatory regions that were found using IVA-GL but not found using ICA-EBM—shown in yellow—correspond to the sensorimotor and parietal regions of the brain. This makes sense, since these are the regions that are expected to be consistently activated across the three tasks. We also see regions, such as some of the auditory activation in the AOD and SM datasets as well as the visual activation in the SIRP, which are found using ICA-EBM, but not using IVA-GL—shown in cyan. This highlights the ability of IVA-GL to emphasize similarities across datasets and also points to the need to compare fusion results with single dataset analyses, since some of the regions associated with a subset of the tasks are not found using IVA-GL.

A primary goal of many neuroimaging studies is the investigation of possible links between cognitive or behavioral measures and neural function, see *e.g.*, [113, 128, 129]. Motivated by this, we examine the relationship between the subject covariations of the GDMs and behavioral scores, which are classified into multiple domains [1], collected from each subject. We display the results in Table 6.2.

From Table 6.2, we see that the GDM found using ICA-EBM on the SIRP dataset does not correlate significantly, after a Bonferroni correction, with the majority of behavioral

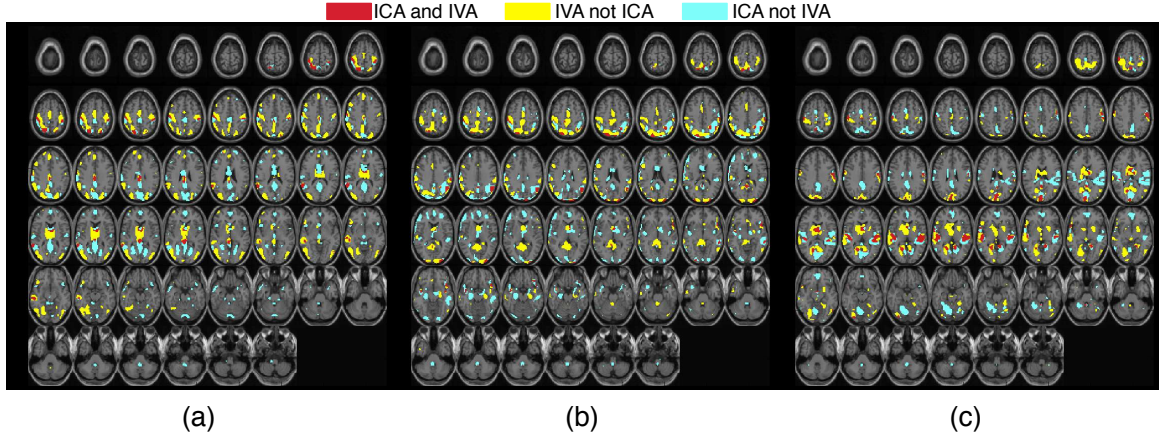


FIG. 6.2. GDMs generated from the (a) AOD, (b) SIRP, and (c) SM datasets. For each plot, we display only those voxels that had an absolute z -score above 1.7. The red shows regions that are common to both the GDM extracted from the IVA-GL results as well as the GDM extracted from the ICA-EBM results. The yellow shows the regions that are in the GDM extracted from IVA-GL results but not in the GDM extracted from the ICA-EBM results. The cyan shows the regions that are in the GDM extracted from ICA-EBM results but not in the GDM extracted from the IVA-GL results.

	t -statistic	AOD (ICA)	AOD (IVA)	SIRP (ICA)	SIRP (IVA)	SM (ICA)	SM (IVA)
Speed of Processing	-5.50	4.25×10^{-7}	2.50×10^{-5}	5.91×10^{-3}	7.86×10^{-6}	1.02×10^{-5}	9.14×10^{-4}
Verbal Learning	11.03	1.06×10^{-6}	6.45×10^{-6}	2.63×10^{-4}	3.03×10^{-6}	4.70×10^{-9}	1.73×10^{-7}
Verbal Working Memory	11.35	5.54×10^{-6}	7.56×10^{-5}	3.58×10^{-3}	7.99×10^{-5}	6.73×10^{-8}	7.10×10^{-5}
Nonverbal Working Memory	7.19	7.04×10^{-4}	1.23×10^{-3}	3.03×10^{-3}	1.95×10^{-5}	1.40×10^{-4}	5.19×10^{-4}

Table 6.2. The first column from the left displays the t -statistic from the two-sample t -test run on the behavioral scores from the “measurement and treatment research to improve cognition in schizophrenia” consensus cognitive battery [1, 2], where the first group is the scores from the controls and the second group is the scores from the patients. The second through seventh columns from the left are the correlation between the subject covariation from the GDM extracted from that combination of dataset and analysis method and the behavioral score. No correction for multiple comparisons was performed, however we highlight the entries that are not significant after a Bonferroni correction— $p < 2.02 \times 10^{-3}$ for a p -value of 0.05—in red.

scores, meaning that it cannot be associated with the results observed clinically. However, the GDM created from the IVA-GL results does correlate significantly. This shows how a joint analysis can provide more clinically informative results than individual analyses, particularly for datasets that are not as discriminative or as well associated with clinical measures, as the SIRP dataset was in this case.

6.3 Summary

Since each BSS technique has different modeling assumptions, applying different techniques to the same datasets will result in different decompositions. This motivates methods to compare these different decompositions and assess their strengths and weaknesses on real data. In this chapter, we propose the use of GDMs to summarize the total discriminative power of each dataset within a decomposition. We apply this technique to multitask fMRI data collected from patients with schizophrenia and healthy controls. Through this application, we find that IVA-GL extracts components that generally represent the common brain regions across the tasks, but can mask regions associated with only a subset of the datasets. We also find that the use of GDMs facilitates comparisons across analysis methods both in terms of the significance of the results and how they are related to certain behavioral scores as well as how the discriminatory brain regions change with the method.

Chapter 7

CONCLUSIONS AND FUTURE WORK

The goal of this dissertation has been the development and analysis of a data-driven fusion framework that can utilize complementary information from different datasets as well as an exploration of the nature and degree of the connections between different datasets. By exploiting both multimodal as well as multiset neuroimaging data, we have sought to provide a better view of spatio-temporal brain function, highlight the similarities across functional tasks, and understand how neural function is disrupted by schizophrenia. We should note that the techniques developed in this dissertation have application far beyond the specific datasets to which they were applied in this dissertation or even the domain that was the focus of our study. Instead, these methods can facilitate understanding of how robust different techniques are to deviations from their assumptions, the effects of including an additional dataset in an analysis, and how different datasets interact to determine a joint result. In this chapter, we summarize our results and present possible directions for further research.

7.1 Conclusions

The extraction of data from multiple sensors enables the exploitation of complementary information from each dataset, thus allowing greater understanding of the problem under study. The datasets can be from sensors of the same type, referred to as multiset

data, such as multitask or multi-subject fMRI data, or from sensors of different types, referred to as multimodal datasets, such as fMRI and EEG data drawn from the same subjects. With the goal of maximizing the amount of complementary information used while minimizing the potential for biases, the performance of data fusion through the use of BSS techniques has recently become quite popular for the analysis of such data. However, the success of these techniques is intimately tied to the problem of order-selection and the degree of similarity, *i.e.*, the amount of complementary information that is shared, among the different datasets. In addition, the classical methods for order-selection only address a single dataset at a time and do not perform well in the sample-poor regime, which is commonly encountered when fusing multimodal neuroimaging datasets using tIVA. In Chapter 3 of this dissertation, we discuss a new method, PCA-CCA, which addresses the problems of order-selection and assessment of the degree of similarity between multimodal neurological datasets in the sample-poor regime. We demonstrate the desirable performance of PCA-CCA compared with classical techniques in a series of simulations. Additionally, we show how the degree of similarity between real multimodal neurological datasets can provide an accurate prediction of the significance of the components generated by tCCA and tMCCA, a topic that we explore more in Chapter 4.

Despite the ability of data fusion to facilitate understanding of neural function and how it is disrupted by disease, most studies focus solely on the fusion of only two modalities. In Chapter 4 of this dissertation, we fuse real multimodal neurological data from *three* datasets, thus enabling exploration of the contribution of each dataset to the fusion result. We find that the number of subjects and the strength of the similarity between the datasets are the primary factors affecting the results achieved by different fusion methods. If the similarity between the datasets is high, such as for the combination of fMRI and ERP datasets, then the more constrained jICA model may be more appropriate than the tIVA

model because the former is less sensitive to additive noise than the latter. However, if the similarity between datasets is not expected to be high, such as for the combination of fMRI and sMRI as well as sMRI and ERP, the tIVA model may be more reasonable, since it is less constrained than the jICA model.

Given that the success of a particular fusion method is dependent on the amount of complementary information that the datasets contain, a question that naturally arises is: what is the most effective method to combine different datasets? In Chapter 5 of this dissertation, we propose a data-driven classification procedure to unambiguously quantify the performance of different analysis techniques as well as the discriminatory information that a new dataset introduces beyond what already exists in other datasets. Through its application to real multitask fMRI data, we find that though we achieve the best performance by analyzing all of the tasks together, the AOD and SM tasks provide much more unique exploitable discriminatory information than the SIRP dataset. We also show that the more flexible IVA-GL algorithm has equivalent or superior performance compared with the popular data fusion method jICA using EBM.

Individual BSS techniques have different modeling assumptions, thus they will produce different decompositions when applied to the same data. In Chapter 6 of this dissertation, we propose a new method, GDMs, to summarize the discriminative power of each dataset within a decomposition. In applying it to real multitask fMRI data, we find that IVA extracts components that generally represent the common brain regions across the tasks, but can mask regions associated with only a subset of the datasets. We also find that the use of GDMs facilitates comparisons across analysis methods both in terms of the significance of the results and how they are related to certain behavioral scores as well as how the discriminatory brain regions change with the method.

7.2 Future Work

The methods developed in this dissertation motivate several potentially fruitful directions for future research.

7.2.1 IVA for Determining the Connections between Multiple Datasets

As we discuss in Chapter 3, the success of PCA-CCA for the determination of the strength of the connections between two multimodal datasets as well as their joint order motivates an extension to more than two datasets. To the best of our knowledge, the topic of determining the number of dependent sources across multiple datasets has received fairly little attention, see *e.g.*, [107, 130–134], and all of these methods exploit solely the second-order dependence, *i.e.*, correlation, among the datasets. Since, as we mention in Chapter 2, IVA is an extension of both CCA and MCCA such that the demixing matrices are not constrained to be orthogonal and higher-order dependencies can be exploited, it provides a natural framework to determine the number of dependent sources across multiple datasets. A simple example that highlights the potential benefits of this approach over previous methods is if there are sources that are uncorrelated but dependent across multiple datasets. In such a case, all previously proposed methods will fail, but an IVA-based technique, which can exploit this higher-order dependence, will succeed.

The IVA-based framework can be combined with two other approaches to increase its applicability to real-world data. First, by assuming a fixed parametric model, such as the generalized Gaussian distribution, for the sources and that the mixing matrices are orthogonal, it may be possible to derive a close-form solution for this IVA problem similar to the case for CCA, MCCA-MAXVAR, and MCCA-MINVAR. Second, bootstrapping techniques, such as the block bootstrap [135], can be used to account for the sample-to-sample dependence encountered in many applications and to generate a null distribution of

values within a hypothesis-testing framework.

7.2.2 ICA+CCA

Recent advances in technology have greatly facilitated the collection of genetic data, such as single nucleotide polymorphisms (SNPs) data or deoxyribonucleic acid methylation data, thus facilitating the discovery of the genetic factors associated with inheritable neurological diseases, such as the early-onset form of Alzheimer’s disease. However, despite the fact that inherited medical and psychiatric diseases are most likely based upon multiple genes, which result in a series of biological processes, as well as environmental factors [136], most genotyping strategies for complex inherited phenotypes perform only univariate analyses, such as genome-wide association studies, see *e.g.*, [137]. More recently, multivariate analyses, of which p-ICA [25] is perhaps the most popular, have become increasingly utilized in order to discover the links between genetic data and neural function. This popularity is likely due to three factors: first, they facilitate interpretation, since they extract functional regions or sets of genes that covary together, second, they provide robustness to noise, and third, because the statistical testing is done at the component level and not based on the SNPs locations—which is usually three to four orders of magnitude higher—thus enabling less extreme corrections for multiple comparisons [138]. Despite this popularity, p-ICA requires the mixing matrices of the two datasets, generally fMRI feature data and SNPs data derived from the same subjects, to be nearly identical; an assumption that may not always be true in practice. As was seen for jICA compared with tIVA in this dissertation, there can be significant benefits to relaxing such restrictive constraints. For this reason, we propose a hybrid technique that combines the strengths of ICA and tCCA, known as ICA+tCCA, in order to factor and then fuse genetic and neurological feature data. Although we discuss the technique below for the case where we have fMRI feature data and SNPs data, the datasets could be any combination of neuroimaging and

genetic datasets, or from other multimodal applications.

Consider an fMRI feature dataset from M subjects, $\mathbf{X}^{[1]} = [\mathbf{x}^{[1]}(1), \dots, \mathbf{x}^{[1]}(V_1)] \in \mathbb{R}^{M \times V_1}$, where the m th row of $\mathbf{X}^{[1]}$ is formed by flattening the feature of V_1 voxels from the m th subject. Similarly, consider a dataset of SNPs collected from the same subjects $\mathbf{X}^{[2]} = [\mathbf{x}^{[2]}(1), \dots, \mathbf{x}^{[2]}(V_2)] \in \mathbb{R}^{M \times V_2}$, where the m th row of $\mathbf{X}^{[2]}$ are the SNPs values from the m th subject. Performing ICA on each dataset separately factorizes them into a set of functional networks $\mathbf{S}^{[1]} \in \mathbb{R}^{C_1 \times V_1}$ and phenotypes $\mathbf{S}^{[2]} \in \mathbb{R}^{C_2 \times V_2}$, with corresponding subject covariations $\mathbf{A}^{[1]} \in \mathbb{R}^{M \times C_1}$ and $\mathbf{A}^{[2]} \in \mathbb{R}^{M \times C_2}$, respectively. Performing tCCA on the subject covariations as described in Chapter 2 will produce combinations of the subject covariations and, therefore, combinations of the factors that maximally correlate across the two datasets. There are several additional positive aspects of this method when compared to p-ICA. First, we can use other factorization methods, such as sparse PCA [139], for either of the two datasets individually or jointly—a feature that is not available for p-ICA. Second, we can extend this technique to analyze data from more than two datasets by replacing the tCCA step with tMCCA. Third, we are not constrained to estimate the same number of factors from each dataset—a strict inherent assumption of p-ICA.

We should also note that this method deviates significantly from the technique CCA+jICA [27, 140]. The primary reason for these differences is do to the fact that the final step of CCA+jICA is jICA. This means that, first, CCA+jICA requires the estimation of the same number of factors from both datasets, and second, CCA+jICA requires the datasets to have nearly the same mixing matrix after the performance of the CCA step in order for the jICA to succeed—an assumption which is not guaranteed in practice.

To conclude, this method also shares some similarities with the technique known as joint and individual variation explained (JIVE) [141]. However, the proposed method depends on ICA to factorize the datasets, whereas JIVE exploits the singular value decompo-

sition (SVD). The superior performance of ICA over second-order methods, such as SVD, for the factorization of individual datasets has been shown consistently, see *e.g.*, [14].

7.2.3 TIVA for Calcium Imaging Data

The goal of studying how the brain encodes information in real-time goal has been greatly advanced by very recent developments in imaging technology, see *e.g.*, [142, 143], which have enabled the study of large populations of neurons in real time using calcium fluorescence imaging. Given the potential offered by such imaging data, many methods have been developed to extract individual neuronal signals from the high levels of noise in the fluorescence images. However, traditional analyses of calcium imaging data have focused on single subject analyses and have compared summary statistics across subjects in order to derive conclusions, see *e.g.*, [144–150]. Data fusion techniques, particularly tIVA, present a unique method to extract coactivating neurons within a task whose timing is identical across subjects, thus facilitating the establishment of how different subjects encode the same stimuli. In addition, since the data is video, *i.e.*, $\mathbf{X}^{[k]} \in \mathbb{R}^{F \times P}$, where P is the number of pixels in the vectorized frame and F is the number of frames in the video, the number of samples for tIVA is F , which is generally quite large. Thus, we can exploit more general models for tIVA than tIVA-G, which are expected to greatly improving performance.

REFERENCES

- [1] K. H. Nuechterlein, M. F. Green, R. S. Kern, L. E. Baade, D. M. Barch, J. D. Cohen, S. Essock, W. S. Fenton, F. J. Frese III, J. M. Gold, T. Goldberg, R. K. Heaton, R. S. E. Keefe, H. Kraemer, R. Mesholam-Gately, L. J. Seidman, E. Stover, D. R. Weinberger, A. S. Young, S. Zalcman, and S. R. Marder, “The MATRICS consensus cognitive battery, part 1: test selection, reliability, and validity,” *American Journal of Psychiatry*, vol. 165, no. 2, pp. 203–213, 2008.
- [2] S. M. August, J. N. Kiwanuka, R. P. McMahon, and J. M. Gold, “The MATRICS consensus cognitive battery (MCCB): clinical and cognitive correlates,” *Schizophrenia Research*, vol. 134, no. 1, pp. 76–82, 2012.
- [3] A. P. James and B. V. Dasarathy, “Medical image fusion: A survey of the state of the art,” *Information Fusion*, vol. 19, pp. 4–19, 2014.
- [4] D. Lahat, T. Adalı, and C. Jutten, “Multimodal data fusion: An overview of methods, challenges, and prospects,” *Proceedings of the IEEE*, vol. 103, pp. 1449–1477, Sep. 2015.
- [5] F. Savopol and C. Armenakis, “Merging of heterogeneous data for emergency mapping: Data integration or data fusion?,” in *International Archives of Photogrammetry Remote Sensing and Spatial Information Sciences*, vol. 34, pp. 668–674, 2002.
- [6] V. D. Calhoun and T. Adalı, “Feature-based fusion of medical imaging data,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, pp. 711–720, Sep. 2009.
- [7] D. Zhang, Y. Wang, L. Zhou, H. Yuan, D. Shen, and Alzheimer’s Disease Neu-

- roimaging Initiative, “Multimodal classification of Alzheimer’s disease and mild cognitive impairment,” *NeuroImage*, vol. 55, no. 3, pp. 856–867, 2011.
- [8] E. Westman, J.-S. Muehlboeck, and A. Simmons, “Combining MRI and CSF measures for classification of Alzheimer’s disease and prediction of mild cognitive impairment conversion,” *NeuroImage*, vol. 62, no. 1, pp. 229–238, 2012.
- [9] C.-Y. Wee, P.-T. Yap, D. Zhang, K. Denny, J. N. Browndyke, G. G. Potter, K. A. Welsh-Bohmer, L. Wang, and D. Shen, “Identification of MCI individuals using structural and functional connectivity networks,” *NeuroImage*, vol. 59, no. 3, pp. 2045–2056, 2012.
- [10] C. Arndt and O. Loffeld, “Information gained by data fusion,” in *Lasers, Optics, and Vision for Productivity in Manufacturing I*, pp. 32–40, International Society for Optics and Photonics, 1996.
- [11] K. Choi, R. C. Craddock, P. E. Holtheimer, Z. Yang, X. Hu, and H. Mayberg, “A combined functional-structural connectivity analysis of major depression using joint independent components analysis,” in *Intl. Soc. Mag. Reson. Med.*, p. 3555, May 2008.
- [12] N. M. Correa, T. Adalı, Y.-O. Li, and V. D. Calhoun, “Canonical correlation analysis for data fusion and group inferences,” *IEEE Signal Processing Magazine*, vol. 27, no. 4, pp. 39–50, 2010.
- [13] T. Adalı, Y. Levin-Schwartz, and V. D. Calhoun, “Multi-modal data fusion using source separation: Two effective models based on ICA and IVA and their properties,” *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1478–1493, 2015.

- [14] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*. John Wiley & Sons, Inc., 2001.
- [15] P. Comon and C. Jutten, *Handbook of Blind Source Separation: Independent Component Analysis and Applications*. Academic Press, 1st ed., 2010.
- [16] V. D. Calhoun, T. Adalı, G. D. Pearlson, and J. J. Pekar, “A method for making group inferences from functional MRI data using independent component analysis,” *Human Brain Mapping*, vol. 14, no. 3, pp. 140–151, 2001.
- [17] M. Svensén, F. Kruggel, and H. Benali, “ICA of fMRI group study data,” *NeuroImage*, vol. 16, no. 3, pp. 551–563, 2002.
- [18] E. Martinez-Montes, P. A. Valdes-Sosa, F. Miwakeichi, R. I. Goldman, and M. S. Cohen, “Concurrent EEG/fMRI analysis by multiway partial least squares,” *NeuroImage*, vol. 22, pp. 1023–1034, 2004.
- [19] V. D. Calhoun, T. Adalı, G. Pearlson, and K. Kiehl, “Neuronal chronometry of target detection: Fusion of hemodynamic and event related potential data,” *NeuroImage*, vol. 30, pp. 544–553, 2006.
- [20] A. Smilde, R. Bro, and P. Geladi, *Multi-way analysis: Applications in the chemical sciences*. John Wiley & Sons, 2005.
- [21] C. F. Beckmann and S. M. Smith, “Tensorial extensions of independent component analysis for multisubject FMRI analysis,” *NeuroImage*, vol. 25, no. 1, pp. 294–311, 2005.
- [22] N. M. Correa, Y.-O. Li, T. Adalı, and V. D. Calhoun, “Canonical correlation analysis for feature-based fusion of biomedical imaging modalities and its application to de-

- tection of associative networks in schizophrenia,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, pp. 998–1007, Dec. 2008.
- [23] J.-H. Lee, T.-W. Lee, F. A. Jolesz, and S.-S. Yoo, “Independent vector analysis (IVA): Multivariate approach for fMRI group study,” *NeuroImage*, vol. 40, no. 1, pp. 86–109, 2008.
- [24] Y. Levin-Schwartz, V. D. Calhoun, and T. Adalı, “Data-driven fusion of EEG, functional and structural MRI: A comparison of two models,” in *48th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, March 2014.
- [25] J. Liu, G. Pearlson, A. Windemuth, G. Ruano, N. I. Perrone-Bizzozero, and V. D. Calhoun, “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.
- [26] J. Sui, T. Adalı, G. D. Pearlson, V. P. Clark, and V. D. Calhoun, “A method for accurate group difference detection by constraining the mixing coefficients in an ICA framework,” *Human Brain Mapping*, vol. 30, no. 9, pp. 2953–2970, 2009.
- [27] J. Sui, T. Adalı, G. Pearlson, H. Yang, S. R. Sponheim, T. White, and V. D. Calhoun, “A CCA+ICA based model for multi-task brain imaging data fusion and its application to schizophrenia,” *NeuroImage*, vol. 51, pp. 123–134, 2010.
- [28] J. Sui, H. He, G. D. Pearlson, T. Adalı, K. A. Kiehl, Q. Yu, V. P. Clark, E. Castro, T. White, B. A. Mueller, B. C. Ho, N. C. Andreasen, and V. D. Calhoun, “Three-way (N-way) fusion of brain imaging data based on mCCA+jICA and its application to discriminating schizophrenia,” *NeuroImage*, vol. 66, pp. 119–132, 2013.

- [29] E. Acar, Y. Levin-Schwartz, V. D. Calhoun, and T. Adalı, “Tensor-based fusion of EEG and fMRI to understand neurological changes in schizophrenia,” in *International Conference on Circuits and Systems (ISCAS)*, 2017, (Accepted).
- [30] T. Kim, T. Eltoft, and T.-W. Lee, “Independent vector analysis: An extension of ICA to multivariate components,” in *Independent Component Analysis and Blind Signal Separation* (J. Rosca, D. Erdogmus, J. Principe, and S. Haykin, eds.), vol. 3889 of *Lecture Notes in Computer Science*, pp. 165–172, Springer Berlin / Heidelberg, 2006.
- [31] V. D. Calhoun, T. Adalı, K. A. Kiehl, R. Astur, J. J. Pekar, and G. D. Pearlson, “A method for multitask fMRI data fusion applied to schizophrenia,” *Human Brain Mapping*, vol. 27, pp. 598–610, 2006.
- [32] M. Ramezani, A. Rasouljan, T. Hollenstein, K. Harkness, I. Johnsrude, and P. Abolmaesumi, “Joint source based analysis of multiple brain structures in studying major depressive disorder,” *Proc. SPIE*, vol. 9034, pp. 1–6, 2014.
- [33] J. Sui, T. Adalı, Q. Yu, J. Chen, and V. D. Calhoun, “A review of multivariate methods for multimodal fusion of brain imaging data,” *Journal of Neuroscience Methods*, vol. 204, no. 1, pp. 68–81, 2012.
- [34] K. A. Kiehl and P. F. Liddle, “An event-related functional magnetic resonance imaging study of an auditory oddball task in schizophrenia,” *Schizophrenia Research*, vol. 48, no. 2, pp. 159–171, 2001.
- [35] R. L. Gollub, J. M. Shoemaker, M. D. King, T. White, S. Ehrlich, S. R. Sponheim, V. P. Clark, J. A. Turner, B. A. Mueller, V. Magnotta, D. O’Leary, B. C. Ho, S. Brauns, D. S. Manoach, L. Seidman, J. R. Bustillo, J. Lauriello, J. Bock-

- polt, K. O. Lim, B. R. Rosen, S. C. Schulz, V. D. Calhoun, and N. C. Andreasen, “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.
- [36] V. P. Clark, S. Fannon, S. Lai, R. Benson, and L. Bauer, “Responses to rare visual target and distractor stimuli using event-related fMRI,” *Journal of Neurophysiology*, vol. 83, no. 5, pp. 3133–3139, 2000.
- [37] V. P. Clark, “A history of randomized task designs in fMRI,” *NeuroImage*, vol. 62, no. 2, pp. 1190–1194, 2012.
- [38] A. M. Michael, S. A. Baum, J. F. Fries, B.-C. Ho, R. K. Pierson, N. C. Andreasen, and V. D. Calhoun, “A method to fuse fMRI tasks through spatial correlations: Applied to schizophrenia,” *Human Brain Mapping*, vol. 30, no. 8, pp. 2512–2529, 2009.
- [39] H. Akaike, “Information theory and an extension of the maximum likelihood principle,” in *Selected Papers of Hirotugu Akaike*, pp. 199–213, Springer, 1998.
- [40] J. Rissanen, “Modeling by shortest data description,” *Automatica*, vol. 14, pp. 465–471, 1978.
- [41] G. Schwarz, “Estimating the dimension of a model,” *Ann. Stat.*, vol. 6, pp. 461–464, 1978.
- [42] X.-L. Li, T. Adalı, and M. Anderson, “Noncircular principal component analysis and its application to model selection,” *IEEE Trans. Signal Process.*, vol. 59, pp. 4516–4528, Oct. 2011.
- [43] M. Wax and T. Kailath, “Detection of signals by information theoretic criteria,” *IEEE Trans. Acoust., Speech and Signal Process.*, vol. 33, pp. 387–392, April 1985.

- [44] Y.-O. Li, T. Adalı, and V. D. Calhoun, “Estimating the number of independent components for functional magnetic resonance imaging data,” *Human Brain Mapping*, vol. 28, no. 11, pp. 1251–1266, 2007.
- [45] P. R. Peres-Neto, D. A. Jackson, and K. M. Somers, “How many principal components? stopping rules for determining the number of non-trivial axes revisited,” *Computational Statistics & Data Analysis*, vol. 49, no. 4, pp. 974–997, 2005.
- [46] H. Yamamoto, T. Fujimori, H. Sato, G. Ishikawa, K. Kami, and Y. Ohashi, “Statistical hypothesis testing of factor loading in principal component analysis and its application to metabolite set enrichment analysis,” *BMC Bioinformatics*, vol. 15, no. 1, pp. 1–9, 2014.
- [47] V. D. Calhoun, T. Adalı, M. C. Stevens, K. A. Kiehl, and J. J. Pekar, “Semi-blind ICA of fMRI: a method for utilizing hypothesis-derived time courses in a spatial ICA analysis,” *NeuroImage*, vol. 25, no. 2, pp. 527–538, 2005.
- [48] C. W. Hesse and C. J. James, “On semi-blind source separation using spatial constraints with applications in EEG analysis,” *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2525–2534, 2006.
- [49] X. Lei, P. A. Valdes-Sosa, and D. Yao, “EEG/fMRI fusion based on independent component analysis: Integration of data-driven and model-driven methods,” *Journal of Integrative Neuroscience*, vol. 11, no. 3, pp. 313–337, 2012.
- [50] C. F. Beckmann, M. Jenkinson, and S. M. Smith, “General multilevel linear modeling for group analysis in FMRI,” *NeuroImage*, vol. 20, no. 2, pp. 1052–1063, 2003.
- [51] P. A. Valdes-Sosa, J. M. Sanchez-Bornot, R. C. Sotero, Y. Iturria-Medina, Y. Aleman-Gomez, J. Bosch-Bayard, F. Carbonell, and T. Ozaki, “Model driven

- EEG/fMRI fusion of brain oscillations,” *Human Brain Mapping*, vol. 30, no. 9, pp. 2701–2721, 2009.
- [52] K. Lee, S. Tak, and J. C. Ye, “A data-driven sparse GLM for fMRI analysis using sparse dictionary learning with MDL criterion,” *IEEE Transactions on Medical Imaging*, vol. 30, pp. 1076–1089, May 2011.
- [53] S. Dähne, F. Bießmann, W. Samek, S. Haufe, D. Goltz, C. Gundlach, A. Villringer, S. Fazli, and K.-R. Müller, “Multivariate machine learning methods for fusing multimodal functional neuroimaging data,” *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1507–1530, 2015.
- [54] D. A. Handwerker, J. M. Ollinger, and M. D’Esposito, “Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses,” *NeuroImage*, vol. 21, no. 4, pp. 1639–1651, 2004.
- [55] V. D. Calhoun, T. Adalı, N. R. Giuliani, J. J. Pekar, K. A. Kiehl, and G. D. Pearlson, “Method for multimodal analysis of independent source differences in schizophrenia: Combining gray matter structural and auditory oddball functional data,” *Human Brain Mapping*, vol. 27, no. 1, pp. 47–62, 2006.
- [56] A. R. Groves, C. F. Beckmann, S. M. Smith, and M. W. Woolrich, “Linked independent component analysis for multimodal data fusion,” *NeuroImage*, vol. 54, no. 3, pp. 2198–2217, 2011.
- [57] W. Swinnen, B. Hunyadi, E. Acar, S. Van Huffe, and M. De Vos, “Incorporating higher dimensionality in joint decomposition of EEG and fMRI,” in *Proceedings of the 22nd European Signal Processing Conference (EUSIPCO)*, pp. 121–125, IEEE, 2014.

- [58] C. Chatzichristos, E. Kofidis, G. Kopsinis, and S. Theodoridis, “Higher-order block term decomposition for spatially folded fMRI data,” *arXiv preprint arXiv:1607.05073*, 2016.
- [59] T. Adalı, M. Anderson, and G.-S. Fu, “Diversity in independent component and vector analyses: Identifiability, algorithms, and applications in medical imaging,” *IEEE Signal Processing Magazine*, vol. 31, pp. 18–33, May 2014.
- [60] A. Bell and T. Sejnowski, “An information maximization approach to blind separation and blind deconvolution,” *Neural Computation*, vol. 7, pp. 1129–1159, 1995.
- [61] X.-L. Li and T. Adalı, “Independent component analysis by entropy bound minimization,” *IEEE Transactions on Signal Processing*, vol. 58, pp. 5151–5164, Dec. 2010.
- [62] V. D. Calhoun, V. K. Potluru, R. Rhlypo, R. F. Silva, B. A. Pearlmutter, A. Caprihan, S. M. Plis, and T. Adalı, “Independent component analysis for brain fMRI does indeed select for maximal independence,” *PLoS ONE*, vol. 8, no. 10, pp. 1–8, 2013.
- [63] T. Adalı, Y. Levin-Schwartz, and V. D. Calhoun, “Multi-modal data fusion using source separation: Application to medical imaging,” *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1494–1506, 2015.
- [64] J.-F. Cardoso, “Infomax and maximum likelihood for blind source separation,” *IEEE Signal Process. Lett.*, vol. 4, pp. 112–114, April 1997.
- [65] M. J. McKeown, S. Makeig, G. G. Brown, T.-P. Jung, S. S. Kindermann, A. J. Bell, and T. J. Sejnowski, “Analysis of fMRI Data by Blind Separation Into Independent Spatial Components,” *Human Brain Mapping*, vol. 6, pp. 160–188, 1998.

- [66] W. Du, Y. Levin-Schwartz, G.-S. Fu, S. Ma, V. D. Calhoun, and T. Adalı, “The role of diversity in complex ICA algorithms for fMRI analysis,” *Journal of Neuroscience Methods*, vol. 264, pp. 129–135, 2016.
- [67] Y. Levin-Schwartz, V. D. Calhoun, and T. Adalı, “Multivariate fusion of EEG and functional MRI data using ICA: algorithm choice and performance analysis,” in *12th International Conference on Latent Variable Analysis and Signal Separation (LVA/ICA)*, pp. 489–496, 2015.
- [68] T. Brookings, S. Ortigue, S. Grafton, and J. Carlson, “Using ICA and realistic BOLD models to obtain joint EEG/fMRI solutions to the problem of source localization,” *NeuroImage*, vol. 44, no. 2, pp. 411–420, 2009.
- [69] K. Specht, R. Zahn, K. Willmes, S. Weis, C. Holtel, B. J. Krause, H. Herzog, and W. Huber, “Joint independent component analysis of structural and functional images reveals complex patterns of functional reorganisation in stroke aphasia,” *NeuroImage*, vol. 47, no. 4, pp. 2057–2063, 2009.
- [70] J. Mangalathu-Arumana, S. A. Beardsley, and E. Liebenthal, “Within-subject joint independent component analysis of simultaneous fMRI/ERP in an auditory oddball paradigm,” *NeuroImage*, vol. 60, no. 4, pp. 2247–2257, 2012.
- [71] Z. Yuan and J. Ye, “Fusion of fNIRS and fMRI data: identifying when and where hemodynamic signals are changing in human brains,” *Frontiers in Human Neuroscience*, vol. 7, pp. 1–9, 2013.
- [72] M. Ramezani, P. Abolmaesumi, K. Marble, H. Trang, and I. Johnsrude, “Fusion analysis of functional MRI data for classification of individuals based on patterns of activation,” *Brain Imaging and Behavior*, vol. 9, no. 2, pp. 149–161, 2014.

- [73] X. Chen, Z. J. Wang, and M. J. McKeown, “A three-step multimodal analysis framework for modeling corticomuscular activity with application to Parkinson’s disease,” *IEEE Journal of Biomedical and Health Informatics*, vol. 18, pp. 1232–1241, July 2014.
- [74] H. Hotelling, “Relations between two sets of variates,” *Biometrika*, vol. 28, no. 3/4, pp. 321–377, 1936.
- [75] J. R. Kettenring, “Canonical analysis of several sets of variables,” *Biometrika*, vol. 58, no. 3, pp. 433–451, 1971.
- [76] M. Borga, O. Friman, P. Lundberg, and H. Knutsson, “A canonical correlation approach to exploratory data analysis in fMRI,” in *Proceedings of the ISMRM Annual Meeting*, (Honolulu, Hawaii), p. 1, May 2002.
- [77] Y.-O. Li, T. Adalı, W. Wang, and V. D. Calhoun, “Joint blind source separation by multiset canonical correlation analysis,” *IEEE Transactions on Signal Processing*, vol. 57, no. 10, pp. 3918–3929, 2009.
- [78] I. Rustandi, M. A. Just, and T. M. Mitchell, “Integrating multiple-study multiple-subject fMRI datasets using canonical correlation analysis,” in *Proc. MICCAI 2009 Workshop: Statist. Model. Detection Issues in Intra- and Inter-Subject Functional MRI Data Anal.*, pp. 1–9, Sep. 2009.
- [79] F. Deleus and M. M. Van Hulle, “Functional connectivity analysis of fMRI data based on regularized multiset canonical correlation analysis,” *Journal of Neuroscience methods*, vol. 197, pp. 143–157, 2011.
- [80] Y. Zhang, G. Zhou, J. Jin, M. Wang, X. Wang, and A. Cichocki, “L1-regularized

- multiway canonical correlation analysis for SSVEP-based BCI,” *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 21, pp. 887–896, Nov. 2013.
- [81] Y.-O. Li, *Novel Second-Order Methods for Data-Driven fMRI Analysis*. PhD thesis, University of Maryland, Baltimore County, 2009.
- [82] A. Michael, M. Anderson, R. Miller, T. Adalı, and V. D. Calhoun, “Preserving subject variability in group fMRI analysis: Performance evaluation of GICA versus IVA,” *Frontiers in Systems Neuroscience*, vol. 8, pp. 106–123, 2014.
- [83] J.-H. Lee, T.-W. Lee, F. A. Jolesz, and S.-S. Yoo, “Independent vector analysis (IVA) for group fMRI processing of subcortical area,” *International Journal of Imaging Systems and Technology*, vol. 18, no. 1, pp. 29–41, 2008.
- [84] M. Anderson, T. Adalı, and X.-L. Li, “Joint blind source separation with multivariate Gaussian model: Algorithms and performance analysis,” *IEEE Transactions on Signal Processing*, vol. 60, pp. 1672–1683, April 2012.
- [85] M. Anderson, G.-S. Fu, R. Phlypo, and T. Adalı, “Independent vector analysis, the Kotz distribution, and performance bounds,” in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3243–3247, 2013.
- [86] J. Himberg, A. Hyvärinen, and F. Esposito, “Validating the independent components of neuroimaging time series via clustering and visualization,” *Neuroimage*, vol. 22, no. 3, pp. 1214–1222, 2004.
- [87] H. W. Kuhn, “The Hungarian method for the assignment problem,” *Naval Research Logistics Quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.
- [88] A. Pezeshki, L. L. Scharf, M. R. Azimi-Sadjadi, and M. Lundberg, “Empirical canonical correlation analysis in subspaces,” in *2004 Conference Record of the*

- Thirty-Eighth Asilomar Conference on Signals, Systems and Computers*, vol. 1, pp. 994–997, Nov. 2004.
- [89] Y. Fujikoshi and L. G. Veitch, “Estimation of dimensionality in canonical correlation analysis,” *Biometrika*, vol. 66, no. 2, pp. 345–351, 1979.
- [90] J. C. Hoefs, “Serum protein concentration and portal pressure determine the ascitic fluid protein concentration in patients with chronic liver disease,” *The Journal of Laboratory and Clinical Medicine*, vol. 102, no. 2, pp. 260–273, 1983.
- [91] J. P. Bush, B. G. Melamed, P. L. Sheras, and P. E. Greenbaum, “Mother–child patterns of coping with anticipatory medical stress,” *Health Psychology*, vol. 5, no. 2, p. 137, 1986.
- [92] G. J. Kennedy, H. R. Kelman, and C. Thomas, “The emergence of depressive symptoms in late life: the importance of declining health and increasing disability,” *Journal of Community Health*, vol. 15, no. 2, pp. 93–104, 1990.
- [93] Z. Lin, C. Zhang, W. Wu, and X. Gao, “Frequency recognition based on canonical correlation analysis for SSVEP-based BCIs,” *IEEE Transactions on Biomedical Engineering*, vol. 53, no. 12, pp. 2610–2614, 2006.
- [94] T. Löfstedt and J. Trygg, “OnPLS - a novel multiblock method for the modelling of predictive and orthogonal variation,” *Journal of Chemometrics*, vol. 25, no. 8, pp. 441–455, 2011.
- [95] O. Alter, P. O. Brown, and D. Botstein, “Generalized singular value decomposition for comparative analysis of genome-scale expression data sets of two different organisms,” *Proceedings of the National Academy of Sciences*, vol. 100, no. 6, pp. 3351–3356, 2003.

- [96] K. van Deun, A. K. Smilde, L. Thorrez, H. A. L. Kiers, and I. van Mechelen, “Identifying common and distinctive processes underlying multiset data,” *Chemometrics and Intelligent Laboratory Systems*, vol. 129, pp. 40–51, 2013.
- [97] Y. Song, P. J. Schreier, D. Ramirez, and T. Hasija, “Canonical correlation analysis of high-dimensional data with very small sample support,” *Signal Processing*, vol. 128, pp. 449–458, 2016.
- [98] R. R. Nadakuditi and A. Edelman, “Sample eigenvalue based detection of high-dimensional signals in white noise using relatively few samples,” *IEEE Transactions on Signal Processing*, vol. 56, no. 7, pp. 2625–2638, 2008.
- [99] W. R. Zwick and W. F. Velicer, “Comparison of five rules for determining the number of components to retain,” *Psychological Bulletin*, vol. 99, no. 3, pp. 432–442, 1986.
- [100] H. Hwang, K. Jung, Y. Takane, and T. S. Woodward, “A unified approach to multiple-set canonical correlation analysis and principal components analysis,” *The British Journal of Mathematical and Statistical Psychology*, vol. 66, no. 2, pp. 308–321, 2013.
- [101] N. J. Roseveare and P. J. Schreier, “Model-order selection for analyzing correlation between two data sets using CCA with PCA preprocessing,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5684–5687, 2015.
- [102] C. Lameiro and P. J. Schreier, “Cross-validation techniques for determining the number of correlated components between two data sets when the number of samples is very small,” in *50th Asilomar Conference on Signals, Systems and Computers*, pp. 601–605, Nov. 2016.

- [103] M. S. Bartlett, "The statistical significance of canonical correlations," *Biometrika*, vol. 32, no. 1, pp. 29–37, 1941.
- [104] D. N. Lawley, "Tests of significance in canonical analysis," *Biometrika*, vol. 46, no. 1/2, pp. 59–66, 1959.
- [105] P. Stoica, K. M. Wong, and Q. Wu, "On a nonparametric detection method for array signal processing in correlated noise fields," *IEEE Transactions on Signal Processing*, vol. 44, pp. 1030–1032, April 1996.
- [106] V. D. Calhoun and T. Adalı, "Multisubject independent component analysis of fMRI: A decade of intrinsic networks, default mode, and neurodiagnostic discovery," *IEEE Reviews in Biomedical Engineering*, vol. 5, pp. 60–73, 2012.
- [107] Y. Wu, K.-W. Tam, and F. Li, "Determination of number of sources with multiple arrays in correlated noise fields," *IEEE Transactions on Signal Processing*, vol. 50, pp. 1257–1260, June 2002.
- [108] E. Pomarol-Clotet, E. J. Canales-Rodríguez, R. Salvador, S. Sarró, J. J. Gomar, F. Vila, J. Ortiz-Gil, Y. Iturria-Medina, A. Capdevila, and P. J. McKenna, "Medial prefrontal cortex pathology in schizophrenia as revealed by convergent findings from multimodal imaging," *Molecular Psychiatry*, vol. 15, no. 8, pp. 823–830, 2010.
- [109] M.-g. Qiu, Z. Ye, Q.-y. Li, G.-j. Liu, B. Xie, and J. Wang, "Changes of brain structure and function in ADHD children," *Brain Topography*, vol. 24, no. 3-4, pp. 243–252, 2011.
- [110] K. Jann, A. Federspiel, S. Giezendanner, J. Andreotti, M. Kottlow, T. Dierks, and T. Koenig, "Linking brain connectivity across different time scales with electroen-

- cephalogram, functional magnetic resonance imaging, and diffusion tensor imaging,” *Brain Connectivity*, vol. 2, no. 1, pp. 11–20, 2012.
- [111] D. Zhang, D. Shen, and Alzheimer’s Disease Neuroimaging Initiative, “Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer’s disease,” *NeuroImage*, vol. 59, no. 2, pp. 895–907, 2012.
- [112] T. M. Schouten, M. Loitfelder, F. de Vos, S. Seiler, J. van der Grond, A. Lechner, A. Hafkemeijer, C. Möller, R. Schmidt, M. de Rooij, and S. A. R. B. Romboutss, “Combining anatomical, diffusion, and resting state functional magnetic resonance imaging for individual classification of mild and moderate Alzheimer’s disease,” *NeuroImage: Clinical*, vol. 11, pp. 46–51, Jan. 2016.
- [113] X. Meng, R. Jiang, D. Lin, J. Bustillo, T. Jones, J. Chen, Q. Yu, Y. Du, Y. Zhang, T. Jiang, J. Sui, and V. D. Calhoun, “Predicting individualized clinical measures by a generalized prediction framework and multimodal fusion of MRI data,” *NeuroImage*, vol. 145, pp. 218–229, 2017.
- [114] J. Sui, R. Huster, Q. Yu, J. M. Segall, and V. D. Calhoun, “Function–structure associations of the brain: evidence from multimodal connectivity and covariance studies,” *NeuroImage*, vol. 102, pp. 11–23, 2014.
- [115] S. Lui, T. Li, W. Deng, L. Jiang, Q. Wu, H. Tang, Q. Yue, X. Huang, R. C. Chan, D. A. Collier, S. A. Meda, G. Pearlson, A. Mechelli, J. A. Sweeney, and Q. Gong, “Short-term effects of antipsychotic treatment on cerebral function in drug-naïve first-episode schizophrenia revealed by resting state functional magnetic resonance imaging,” *Archives of General Psychiatry*, vol. 67, no. 8, pp. 783–792, 2010.
- [116] B.-C. Ho, N. C. Andreasen, S. Ziebell, R. Pierson, and V. Magnotta, “Long-term

- antipsychotic treatment and brain volumes: a longitudinal study of first-episode schizophrenia,” *Archives of General Psychiatry*, vol. 68, no. 2, pp. 128–137, 2011.
- [117] S. C. Strother, J. Anderson, L. K. Hansen, U. Kjems, R. Kustra, J. Sidtis, S. Frutiger, S. Muley, S. LaConte, and D. Rottenberg, “The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework,” *NeuroImage*, vol. 15, no. 4, pp. 747–771, 2002.
- [118] S. Ma, V. D. Calhoun, R. Phlypo, and T. Adalı, “Dynamic changes of spatial functional network connectivity in healthy individuals and schizophrenia patients using independent vector analysis,” *NeuroImage*, vol. 90, pp. 196–206, 2014.
- [119] S. Gopal, R. L. Miller, A. Michael, T. Adalı, M. Cetin, S. Rachakonda, J. R. Bustillo, N. Cahill, S. A. Baum, and V. D. Calhoun, “Spatial variance in resting fMRI networks of schizophrenia patients: an independent vector analysis,” *Schizophrenia Bulletin*, vol. 42, no. 1, pp. 152–160, 2016.
- [120] D. I. Kim, D. H. Mathalon, J. M. Ford, M. Mannell, J. A. Turner, G. G. Brown, A. Belger, R. Gollub, J. Lauriello, C. Wible, D. OLeary, K. Lim, A. Toga, S. G. Potkin, F. Birn, and V. D. Calhoun, “Auditory oddball deficits in schizophrenia: an independent component analysis of the fMRI multisite function BIRN study,” *Schizophrenia Bulletin*, vol. 35, no. 1, pp. 67–81, 2009.
- [121] F. Schneider, U. Habel, M. Reske, T. Kellermann, T. Stöcker, N. J. Shah, K. Zilles, D. F. Braus, A. Schmitt, R. Schlösser, M. Wagner, I. Frommann, T. Kircher, A. Rapp, E. Meisenzahl, S. Ufer, S. Ruhrmann, R. Thienel, H. Sauer, F. A. Henn, and W. Gaebel, “Neural correlates of working memory dysfunction in first-episode schizophrenia patients: an fMRI multi-center study,” *Schizophrenia Research*, vol. 89, no. 1, pp. 198–210, 2007.

- [122] L. Deserno, P. Sterzer, T. Wüstenberg, A. Heinz, and F. Schlagenhauf, “Reduced prefrontal-parietal effective connectivity and working memory deficits in schizophrenia,” *Journal of Neuroscience*, vol. 32, no. 1, pp. 12–20, 2012.
- [123] R. E. Gur, B. I. Turetsky, J. Loughead, W. Snyder, C. Kohler, M. Elliott, R. Pratiwadi, J. D. Ragland, W. B. Bilker, S. J. Siegel, S. J. Kanes, S. E. Arnold, and R. C. Gur, “Visual attention circuitry in schizophrenia investigated with oddball event-related functional magnetic resonance imaging,” *American Journal of Psychiatry*, vol. 164, no. 3, pp. 442–449, 2007.
- [124] S. Whitfield-Gabrieli, H. W. Thermenos, S. Milanovic, M. T. Tsuang, S. V. Faraone, R. W. McCarley, M. E. Shenton, A. I. Green, A. Nieto-Castanon, P. LaViolette, J. Wojcikb, J. D. E. Gabrielia, and L. J. Seidman, “Hyperactivity and hyperconnectivity of the default network in schizophrenia and in first-degree relatives of persons with schizophrenia,” *Proceedings of the National Academy of Sciences*, vol. 106, no. 4, pp. 1279–1284, 2009.
- [125] L. L. Sears, N. C. Andreasen, and D. S. O’leary, “Cerebellar functional abnormalities in schizophrenia are suggested by classical eyeblink conditioning,” *Biological Psychiatry*, vol. 48, no. 3, pp. 204–209, 2000.
- [126] D. Zhang, D. Shen, and Alzheimer’s Disease Neuroimaging Initiative, “Predicting future clinical changes of MCI patients using longitudinal and multimodal biomarkers,” *PloS One*, vol. 7, no. 3, pp. 1–15, 2012.
- [127] M. Ramezani, K. Marble, H. Trang, I. S. Johnsrude, and P. Abolmaesumi, “Joint sparse representation of brain activity patterns in multi-task fMRI data,” *IEEE Transactions on Medical Imaging*, vol. 34, no. 1, pp. 2–12, 2015.

- [128] L. Shen, P. M. Thompson, S. G. Potkin, L. Bertram, L. A. Farrer, T. M. Foroud, R. C. Green, X. Hu, M. J. Huentelman, S. Kim, J. S. K. Kauwe, Q. Li, E. Liu, F. Macciardi, J. H. Moore, L. Munsie, K. Nho, V. K. Ramanan, S. L. Risacher, S. Stone, David J. Swaminathan, A. W. Toga, M. W. Weiner, and A. J. Saykin, “Genetic analysis of quantitative phenotypes in AD and MCI: imaging, cognition and biomarkers,” *Brain Imaging and Behavior*, vol. 8, no. 2, pp. 183–207, 2014.
- [129] J. Wan, Z. Zhang, B. D. Rao, S. Fang, J. Yan, A. J. Saykin, and L. Shen, “Identifying the neuroanatomical basis of cognitive impairment in Alzheimer’s disease by correlation-and nonlinearity-aware sparse Bayesian learning,” *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1475–1487, 2014.
- [130] M. Bhandary, “Detection of the number of signals in the presence of white noise in decentralized processing,” *IEEE Transactions on Signal Processing*, vol. 46, no. 3, pp. 800–803, 1998.
- [131] T. Hasija, Y. Song, P. J. Schreier, and D. Ramírez, “Detecting the dimension of the subspace correlated across multiple data sets in the sample poor regime,” in *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pp. 1–5, June 2016.
- [132] Y. Song, T. Hasija, P. J. Schreier, and D. Ramírez, “Determining the number of signals correlated across multiple data sets for small sample support,” in *24th European Signal Processing Conference (EUSIPCO)*, pp. 1528–1532, IEEE, 2016.
- [133] T. Hasija, Y. Song, P. J. Schreier, and D. Ramirez, “Bootstrap-based detection of the number of signals correlated across multiple data sets,” in *50th Asilomar Conference on Signals, Systems and Computers*, pp. 610–614, Nov. 2016.
- [134] S. Bhinge, Y. Levin-Schwartz, and T. Adalı, “Estimation of common subspace order

- across multiple datasets: Application to multi-subject fMRI data,” in *51st Annual Conference on Information Sciences and Systems (CISS)*, March 2017.
- [135] S. N. Lahiri, *Resampling methods for dependent data*. Springer Science & Business Media, 2013.
- [136] M. J. Owen, N. Craddock, and M. C. ODonovan, “Suggestion of roles for both common and rare risk variants in genome-wide studies of schizophrenia,” *Archives of General Psychiatry*, vol. 67, no. 7, pp. 667–673, 2010.
- [137] M. I. McCarthy, G. R. Abecasis, L. R. Cardon, D. B. Goldstein, J. Little, J. P. A. Ioannidis, and J. N. Hirschhorn, “Genome-wide association studies for complex traits: consensus, uncertainty and challenges,” *Nature Reviews Genetics*, vol. 9, no. 5, pp. 356–369, 2008.
- [138] G. D. Pearlson, V. D. Calhoun, and J. Liu, “An introductory review of parallel independent component analysis (p-ICA) and a guide to applying p-ICA to genetic data and imaging phenotypes to identify disease-associated biological pathways and systems in common complex disorders,” *Frontiers in Genetics*, vol. 6, no. 276, pp. 1–13, 2015.
- [139] H. Zou, T. Hastie, and R. Tibshirani, “Sparse principal component analysis,” *Journal of Computational and Graphical Statistics*, vol. 15, no. 2, pp. 265–286, 2006.
- [140] Y.-O. Li, T. Adalı, and V. D. Calhoun, “A multivariate model for comparison of two datasets and its application to fMRI analysis,” in *2007 IEEE Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 217–222, IEEE, 2007.
- [141] E. F. Lock, K. A. Hoadley, J. S. Marron, and A. B. Nobel, “Joint and individual

variation explained (JIVE) for integrated analysis of multiple data types,” *The Annals of Applied Statistics*, vol. 7, no. 1, p. 523, 2013.

- [142] K. K. Ghosh, L. D. Burns, E. D. Cocker, A. Nimmerjahn, Y. Ziv, A. El Gamal, and M. J. Schnitzer, “Miniaturized integration of a fluorescence microscope,” *Nature Methods*, vol. 8, no. 10, pp. 871–878, 2011.
- [143] Y. Ziv and K. K. Ghosh, “Miniature microscopes for large-scale imaging of neuronal activity in freely behaving rodents,” *Current Opinion in Neurobiology*, vol. 32, pp. 141–147, 2015.
- [144] C. M. Niell and S. J. Smith, “Functional imaging reveals rapid development of visual response properties in the zebrafish tectum,” *Neuron*, vol. 45, no. 6, pp. 941–951, 2005.
- [145] E. Yaksi and R. W. Friedrich, “Reconstruction of firing rate changes across neuronal populations by temporally deconvolved Ca^{2+} imaging,” *Nature Methods*, vol. 3, no. 5, pp. 377–383, 2006.
- [146] E. A. Mukamel, A. Nimmerjahn, and M. J. Schnitzer, “Automated analysis of cellular signals from large-scale calcium imaging data,” *Neuron*, vol. 63, no. 6, pp. 747–760, 2009.
- [147] M. Pachitariu, A. M. Packer, N. Pettit, H. Dalgleish, M. Hausser, and M. Sahani, “Extracting regions of interest from biological images with convolutional sparse block coding,” in *Advances in Neural Information Processing Systems*, pp. 1745–1753, 2013.
- [148] J. Tomek, O. Novak, and J. Syka, “Two-photon processor and SeNeCA: a freely available software package to process data from two-photon calcium imaging

at speeds down to several milliseconds per frame,” *Journal of Neurophysiology*, vol. 110, no. 1, pp. 243–256, 2013.

- [149] R. Maruyama, K. Maeda, H. Moroda, I. Kato, M. Inoue, H. Miyakawa, and T. Aonishi, “Detecting cells using non-negative matrix factorization on calcium imaging data,” *Neural Networks*, vol. 55, pp. 11–19, 2014.
- [150] E. A. Pnevmatikakis, D. Soudry, Y. Gao, T. A. Machado, J. Merel, D. Pfau, T. Reardon, Y. Mu, C. Lacefield, W. Yang, M. Ahrens, R. Bruno, T. M. Jessell, D. S. Peterka, R. Yuste, and L. Paninski, “Simultaneous denoising, deconvolution, and demixing of calcium imaging data,” *Neuron*, vol. 89, no. 2, pp. 285–299, 2016.

