Multi-view Representation Learning with Applications to Functional Neuroimaging Data

Po-Hsuan Chen

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Electrical Engineering
Adviser: Professor Peter J. Ramadge

September 2017

Abstract

One of the greatest challenges for the 21st century is understanding how the human brain works. Although there are different levels of understanding of the human brain, a key step is knowing how brain activity patterns map onto cognition, emotion, memories, etc. This can be studied using functional magnetic resonance imaging (fMRI). fMRI is a non-invasive brain imaging technique with unprecedented spatiotemporal resolution. The fMRI data is gathered while subjects perform a wide-range of cognitive tasks. Analysis of fMRI data using multivariate statistics and machine learning has led to tremendous success in understanding how patterns of neural activity reflect mental representations. This thesis aims to continue the success through advancing machine learning methods motivated by applications to neuroscience problems.

We develop a multi-view learning framework that estimates shared features from multi-view data. We analyze and demonstrate two primary approaches of how can a multi-view learning framework provide new ways of exploring neuroimaging data. First, a multi-view learning model forms a larger dataset by aggregating data from multiple views. A key potential advantage of this is an increase in statistical sensitivity. Second, a multi-view learning model learns a shared feature space and transformations between each view's observation space and the shared feature space. These transformations bridge any two views, opening up new possibilities for analyzing the data. For example, by treating a subject as a view, we can transform one subject's fMRI data into the space of another subject's brain. By treating semantic vectors of stimulus text description and fMRI response as different views, it opens up the opportunity to generate text from fMRI responses or fMRI responses from text.

Lastly, we explore various forms of multi-view learning models, including manifold learning, probabilistic modeling, deep neural network, etc. Different ways of applying multi-view models on neuroimaging data are demonstrated and analyzed. We also discuss our contribution to the open-source software community.

Acknowledgements

Ph.D. studies is a long but extremely rewarding journey for me. During the past five years, I have learned and grown so much in research and various other aspects of life, making me a much more independent researcher as well as a much more mature person. This thesis is one of the manifestations of the journey, and its completion would not have been possible without the support from a large group of people.

First and foremost, I would like to express my highest appreciation to my adviser, Prof. Peter Ramadge. I owe a great debt of gratitude to him for his tremendous support and guidance throughout this journey. He provided me with much freedom to explore issues and directions that interest me and always offered me extremely useful and thoughtful feedback during the exploration. Through the past five years, I learned from him how to be truthful when encountering a problem and how to solve a problem with a down-to-earth approach. I have always enjoyed our weekly meetings, and this is definitely what I'll miss the most after leaving Princeton.

I would also like to thank Prof. Uri Hasson for the numerous discussions we have had on research. While we are pushing for the advancement in machine learning methods, he always reminds me of the scientific questions we are trying to answer. This keeps me from getting lost in the technical details and remaining focused on learning more about the brain.

I will not be able to reach the current stage without the support from several committees that I have worked with along the way. I would like to thank Prof. Janice Chen and Prof. Yuxin Chen for their thorough reviews of the thesis. The insights and suggestions they provided are very valuable in improving the thesis quality. I would also like to thank Prof. Uri Hasson and Prof. Ken Norman for being on my FPO committee. I have always enjoyed discussing my research with them. Lastly, I would like to thank my general examination committee, Prof. David Blei, and Prof. Paul Cuff. A few years ago, when some of my works were still in a very early stage, I

had the opportunity to discuss those with them, and the feedback they provided has motivated the developments of several sections in the thesis.

It has been a privilege and a great honor to work with lots of world-class researchers from many excellent research groups in the past five years. First of all, I would like to thank everyone in the Ramadge lab, including Hejia Zhang, Hossein Valavi, Xu (Mia) Chen, Yun Wang, David Eis, Hao Xu, Pingmei Xu, Alex Lorbert, for their support in research. Special thanks go to Selina Man, holiday parties at your house with all the amazing food and enthusiastic chat are one of my favorite events at Princeton. I would also like to thank everyone in the PNI-Intel project for all the helpful discussions, comments, and feedback on my research. Most of my research is rooted from this big collaborative project, and I have learned so much from all the group members. I really enjoyed the discussions we have had; individually, as small groups, or in the weekly Friday meeting. This is a very long list of the group members, who have all provided either guidance or support for me, including Prof. Jonathan D Cohen, Prof. Kenneth A Norman, Prof. Nicholas B Turk-Browne, Prof. Jonathan Pillow, Prof. Janice Chen, Prof. Jeremy R. Manning, J. Benjamin Hutchinson, Christopher A. Baldassano, Edwin C. Clayton, Mingbo Cai, Michael Shvartsman, Yida Wang, Yaara Yeshurun, Michael J. Arcaro, Kiran Vodrahalli, Sebastian Musslick, Anqi Wu, Theodore L Willke, Javier Turek, Xia (Ivy) Zhu, and Mihai Capota. Last but not the least, I would like to thank another group of amazing neuroscientists, with whom I had the great opportunity to work with. They are members of the Haxby lab, including Prof. James Haxby, Prof. Michael Hanke, Prof. Yaroslav O. Halchenko, and J. Swaroop Guntupalli.

Much thanks to the staff in the Princeton Electrical Engineering Department, Colleen Conrad, Sarah McGovern, Roelie Abdi-Stoffers, Katharine Furda, Lisa Lewis, and Dorothy Coakley, and the staff in the Princeton Neuroscience Institute, Benjamin Singer, PNI help desk, and Alexandra M. Michaud, for their enormous help.

This dissertation would not have been possible without the funding support from Google PhD Fellowship, Princeton University Fellowship in Natural Science and Engineering, Taiwan Ministry of Education Study Abroad Scholarship, Intel, and NSF.

I would not be able to overcome all the challenges along the long journey without the support from my friends. I am beyond blessed to have so many amazing friends, who are like my family. I would like to thank all my friends scattered around the world for listening, offering me advice, encouraging me, and supporting me through the entire process. Whenever I need someone, they are always there for me. I appreciate all the messages, emails, phone calls, and visits. Cheers to many more years of laughter and friendship to come!

To whoever is reading this, I would like to thank you for your time and I sincerely hope that you will get something out of this thesis!

Finally and most importantly, I would like to thank my dad, my mom, and my brother, Chin-Chyuan Chen, Shu-Chao Wu, and Po-Heng Chen. Life is full of ups and downs, but they are always my strongest support and my greatest motivation. The unconditional love that they provided is always the largest comfort for me. Without them, I will not be the person I am today. This thesis is dedicated to them!

To my family.

Contents

Abstract iii

Acknowledgements iv

List of Tables xi

List of Figures xii

1 Introduction **1**

1.1 What is Multi-view Representation Learning? 1

1.2 What is Neuroimaging? 3

1.3 The Need for Multi-view Representation Learning in Neuroimaging . 4

1.4 Organization and Contribution of the Thesis 5

2 A Shared Response Model **8**

2.1 Introduction 8

2.2 Shared Response Framework 10

2.3 Probabilistic Shared Response Model (SRM) 17

2.4 Parameter Estimation for SRM 19

2.5 Connections with Related Methods 21

2.6 Empirical Test of SRM with Synthetic Data 28

2.7 Discussion and Conclusion 29

3 Shared Response Model on Neuroimaging Data **31**

3.1 Introduction 31

3.2 fMRI Datasets 34
3.3 SRM and Spatial Smoothing 36
3.4 Temporal Similarity and Image Classification 38
3.5 Differentiating between Groups 41
3.6 Retinotopy Data 45
3.7 Searchlight Shared Response Model 48
3.8 Amount of Data Required for SRM 53
3.9 Other Explorations 56
3.10 Discussion and Conclusion 57

4 Extensions of SRM **60**
4.1 Introduction 60
4.2 Semi-supervised Shared Response Model 61
4.3 Shared Response Independent Component Analysis 65
4.4 Kernelized Shared Response Model 73
4.5 Gaussian Process Shared Response Model 78

5 A Multi-view Convolutional Autoencoder **84**
5.1 Introduction 84
5.2 Limitations of Current Methods 88
5.3 Fully-connected Autoencoder and SRM 89
5.4 A Multi-view Convolutional Autoencoder 92
5.5 Experiments and Results 95
5.6 Discussion and Conclusion 101

6 Beyond Multiclass Classification **104**
6.1 Introduction 104
6.2 Bridging Semantic Space and fMRI Space 107
6.3 Experiments 109

6.4 Discussion and Conclusion 111

7 Conclusion **113**

A Prior Presentations and Publications **116**

A.1 Prior Presentations 116

A.2 Prior Publications 118

Bibliography **120**

List of Tables

3.1 List of fMRI datasets 36

4.1 Comparison of average accuracy for brain decoding experiments. . . . 64

4.2 Comparison of computation and memory complexity. 82

5.1 List of fMRI datasets 94

List of Figures

2.1 Comparison of two problem formulations (2.2) and (2.1) 11

2.2 Illustration of shared response framework 13

2.3 Dropping orthogonality 17

2.4 Graphical model for SRM 19

2.5 SRM on synthetic dataset 29

3.1 SRM and spatial smoothing 37

3.2 Experiment diagrams for time segment matching and image classification 40

3.3 Experiment results for time segment matching and image classification 41

3.4 Differentiating between groups experiment 42

3.5 SRM with retinotopy data 46

3.6 Between subject transformation of retinotopy with SRM 47

3.7 Time segment matching of sherlock movie dataset with searchlight SRM 49

3.8 Time segment matching of audiobook dataset with searchlight SRM . 51

3.9 Recall matching of sherlock-recall dataset with searchlight SRM . . . 52

3.10 Time segment and scene recall matching with various algorithms . . . 53

3.11 Evaluating SRM with different number of subjects 54

3.12 Evaluating SRM with different number of TRs 55

3.13 SRM with non-temporally synchronized stimulus 56

4.1 Average accuracy as a function of the number of training samples . . 64

4.2 Accuracy and k maps for time segment matching on *sherlock-movie* . 69
4.3 Accuracy and k brain maps for time segment matching with *audiobook* 71
4.4 Accuracy and k brain maps for scene recall matching with *sherlock-recall* 72
4.5 Time segment matching and scene recall matching accuracies 73

5.1 Illustration of lack of spatial locality 87
5.2 Tied-weights linear fully-connected multi-view autoencoder 90
5.3 A nonlinear fully-connected multi-view autoencoder 90
5.4 4D convolutional autoencoder (CAE) 92
5.5 Accuracy maps for time segment and scene recall matching with CAE 97
5.6 Accuracy for whole brain time segment and scene recall matching . . 99
5.7 Dispersion experiment comparison 100

6.1 Bridging semantic space and fMRI response space 107
6.2 Visualization of DMN, vLANG, and dLANG ROIs 109
6.3 Bidirectional accuracy on scene classification and ranking 111

Chapter 1

Introduction

1.1 What is Multi-view Representation Learning?

Multi-view representation learning is a family of methods for learning an unknown underlying representation using a dataset that is multi-view in nature. The key outputs from multi-view learning are an unknown underlying representation, and the mapping from the input space of each view to the common space. Here we use the term "multi-view" in a very broad sense; it can take various specific forms, such as multi-subject (different people), multi-modality (different types of neuroimaging technique), multi-region (different parts of the brain), multi-manifestation (image+text, audio+video, parallel text in different languages), etc. The key notion is that there's something shared or common across different views. Traditional machine learning approaches ignore the fact that data comes in different views, and simply treats a multi-view dataset as a single-view dataset. For example, typical fMRI data analysis simply conducts anatomical registration across different subjects and then analyzes in the anatomical template space. Successful multi-view learning can lead to better utilization of the data yielding better predictive performance.

The concept of multi-view representation learning has become increasingly prominent these days. However, there is not yet a precise definition that circumscribes the core idea of multi-view representation learning. One of the earlier and classical multi-view models is Canonical Correlation Analysis (CCA) [63]. This learns view specific linear transformations such that the correlation between two transformed datasets is maximized. CCA has been useful in various applications [56]. For example, computer vision [72], time series[4] , genomic data [133, 98], etc. CCA has also been extended to a probabilistic formulation [13], a nonlinear formulation [3, 89, 12], and a sparse formulation [9, 30, 55]. The classical CCA is limited to data from two views. In this case, it can be solved as a generalized eigenvalue problem [107]. A generalization of classical CCA, extending two views to multi-view has been proposed in [71], which is called multi-set CCA (MCCA). MCCA turns out to be a much more difficult problem. It has been shown to be NP-hard [107]. MCCA has been widely applied in various problems, including neuroimaging [81], speech recognition [10], word embedding [35], clustering [24], etc.

Besides CCA-like models, there are many other models that fall into multi-view learning framework. For applications in neuroimaging, for example, several different families of multi-view learning models besides maximizing correlation (CCA like) have been developed, including dictionary learning [122, 1, 121, 38, 90, 36], maximization variance (PCA like) [60, 134, 83, 28, 50, 22], maximizing independence (ICA like)[20, 80, 21, 123, 40, 91, 78, 22], structural factor analysis [85, 84], etc. Furthermore, recently, there have been rapid developments in multi-view learning with deep neural network, including fundamental model development [8, 127, 128] to various applications spanning recommendation system [39], image caption generation [42, 86], language model [74], computer vision[95, 136], etc.

Multi-view learning in some sense is also similar to the machine learning notions of transfer learning or multi-task learning. These terminologies generally refer to

accelerated learning of a new task or better generalization to a new task through related tasks that were previously learned or are learned in parallel. In multi-view learning, by learning a common representation across views, we can utilize data from other views to analyze the data from the "testing view" we are interested in.

1.2 What is Neuroimaging?

One of the greatest challenges for the 21st century is understanding how the human brain works. The human brain is the source of memories, emotions, and thoughts. A better understanding of the human brain facilitates the development of human society in enormous ways, including, but not limited to science, medicine, and education. There are different levels of understanding the human brain, but a key step is knowing how brain activity patterns support cognition, emotion, memories, etc. To build these connections, we need to measure brain activity. Neuroimaging is a collection of techniques and methods for measuring brain activity. This includes various techniques to image the structure and the functional response of the brain. There are many types of neuroimaging modalities, including computed tomography (CT), diffuse optical image (DOI), event-related optical signal (EROS), magnetic resonance imaging (MRI), magnetoencephalography (MEG), positron emission tomography (PET), single-photon emission computed tomography (SPECT), near-infrared spectroscopy (NIRS), electroencephalography (EEG), electrocorticography (ECoG), etc. Among various types of neuroimaging data, this thesis will focus on functional Magnetic Resonance Imaging (fMRI) [64]. fMRI is a non-invasive imaging technique measuring neural activity by using the blood-oxygen-level dependent (BOLD) [64] contrast as a proxy for neural activation [82]. It measures human brain activity through oxygenated blood. Among non-invasive brain imaging techniques, fMRI has

unprecedented spatiotemporal resolution with no known side effects. The data is gathered while subjects perform a wide-range of cognitive tasks.

1.3 The Need for Multi-view Representation Learning in Neuroimaging

Over the past 15 years, application of multivariate statistics and machine learning to fMRI data has lead to tremendous success in understanding how patterns of neural activity code mental representations [59, 97, 92, 102, 31]. We argue that multi-view learning methods will play a key role in continuing this success.

First of all, a key component of future fMRI research will be the use of multi-subject datasets. Fundamentally, the use of multi-subject data is critical for assessing the generality and validity of the findings across subjects. Furthermore, one can gather at most a few thousand noisy instances of fMRI data from a single subject. Aggregating multi-subject data to form a larger dataset is essential for increasing the power of multivariate statistic analysis. The assumption here is that there is information shared across subjects [58], and aggregating data means extracting this shared information. If there's nothing shared across subjects, there's little hope for extracting more information by the use of multiple subjects. Shared information could be of various forms. It could be a set of spatial response patterns, latent time-series of the underlying neural activity, or even networks of how various parts of the brain work together. Extracting shared information across subjects requires resolving a major problem. Both anatomical structure and functional topography (patterns of brain activity) vary across subjects [114, 129, 118, 88]. There are existing methods of anatomical alignment [114, 88, 44] , however, it is well known that standard methods of anatomical alignment [114, 88, 44] do not adequately align functional topography [88, 18, 108, 32, 33]. Simply averaging subjects' fMRI responses (captured under a

shared time-synchronized stimulus) after standard anatomical alignment[114, 88, 44], is insufficient to address the variability in functional topography [108, 32, 33]. Indeed, even if "perfect" anatomical alignment was possible, it would not align functional topography [88, 18, 108, 32, 33]. Hence anatomical alignment is often followed by spatial smoothing of the data to blur functional topographies. This is not ideal, and this is where multi-view learning comes into play.

Another key component to understanding how patterns of neural activity code mental representations is the development of encoding and decoding models [93] for neuroimaging data. For encoding, we are interested in predicting brain activity from stimuli, while for decoding, we use brain activity to predict stimuli. Multi-view representation learning plays a key role in bridging between stimuli and brain activity by treating them as different views of the same underlying representation. We can directly model brain activity from stimulus or model stimulus from brain activity. However, in the multi-view learning framework, we will be jointly learning a common representation across brain activity and stimulus. Through the common representation, we can also arbitrarily map brain activity into stimulus or the other way around.

1.4 Organization and Contribution of the Thesis

In Chapter 2, we introduce a framework in a factor model form to identify a shared response. We call this a shared response framework. We first present the shared response framework as simply a multi-view factor model. By analyzing a different formulation, we then proposed a probabilistic shared response model (SRM). Connections with related methods are drawn, and an experiment with synthetic data is conducted and discussed at the end of the chapter.

In Chapter 3, we introduce how we can use SRM with neuroimaging data. This chapter demonstrates the usage of the SRM on fMRI datasets. In addition it gives a reference guide for neuroscientists interested in exploring how can they use SRM to aid their research. We conduct a series of experiments to verify two things. First, SRM's generalizability to a new stimulus, new subjects, and new areas. Second, SRM's effectiveness in decoupling shared and individual response from a group of subjects.

In Chapter 4, we extend SRM by incorporating various additional desired properties. We start with the idea of combining both labeled and unlabeled data to learn a better shared feature space. This leads to a semi-supervised extension of SRM [120]. Independent component SRM replaces the default SRM objective function with an alternative objective function that's similar to Independent Component Analysis [67]. Kernelized Shared Response Model (KSRM) extends standard SRM by the introduction of kernel, and Gaussian process SRM imposes temporal structure on the latent factor.

In Chapter 5, we propose a multi-view convolutional autoencoder for multi-subject fMRI data. The network preserves spatial locality when aggregating information across different subjects using convolutional filters. The aim is to improve anatomical and functional interpretability of the analysis results.

The previous chapters evaluate model performance using some form of prediction. This can also be viewed as hypothesis testing. In Chapter 6, we go beyond the prediction framework by discussing encoding and decoding models. With an encoding model, we map stimulus features to brain responses. With an decoding model, we map brain response directly to stimulus features. This opens up the possibility to decode brain responses directly to text, going beyond multiclass classification.

In Chapter 7, we then draw conclusions and review the contribution of the thesis. Besides the theoretical analysis and empirical evaluation of methods and algorithms

in the thesis, we also make our research easily reproducible and open to the public by developing and contributing to open-source software using publicly available dataset.

Prior Publications Parts of this thesis have been published in [27, 29, 139, 126, 120, 7, 28].

Chapter 2

A Shared Response Model

2.1 Introduction

A key aspect of using multi-view data is properly aggregating the data across views. For example, in neuroscience research, an underlying hypothesis is that under a shared stimulus there is information shared across the subjects' fMRI responses [58]. Aggregating the fMRI data means extracting this shared information[1]. Shared information could be of various forms. For example, in fMRI, it could be a set of spatial response patterns, latent time-series of the underlying neural activity, or networks indicating how various parts of the brain work together.

In this chapter, we develop a framework that models the shared response as a latent variable, and focus on estimating this latent variable through multi-view data. We first focus on developing a general framework and explore its mathematical properties. We then move on to the application of the SRM to fMRI data in Chapter 3. Various extensions of SRM motivated by the application to fMRI data are explored in later chapters.

[1]The term information is used in a generic sense, not necessary the *information* in information theory.

Multivariate statistical analysis often begins by identifying a set of features that capture the informative aspects of the data. For example, one might select a subset of hand-crafted features, or select a subset of principal components of the data. Then use these features for subsequent analysis.

In a similar way, one can think of the multi-view problem as a two-step process. First, we use training data to learn a mapping of each view's data to a shared feature space in a way that captures the across-view shared response. Then use these learned mappings to project held out data for each view into the shared feature space and perform a statistical analysis.

To make this more precise, let $\{X_i \in \mathbb{R}^{v \times d}\}_{i=1}^{m}$ denote matrices of training data (v dimension of observation, over d time points) for m views. We propose using this data to learn view specific basis $W_i \in \mathbb{R}^{v \times k}$, where k is to be selected, and a shared matrix $S \in \mathbb{R}^{k \times d}$ of feature responses such that $X_i = W_i S + E_i$ where E_i is an error term corresponding to unmodeled aspects of the view's data. One can think of W_i as representing a view specific basis and S as a latent feature that captures the component of the response shared across views. We don't claim that S is a sufficient statistic, but that is a useful analogy.

The contribution of this chapter is twofold: First, we propose a general framework for identifying the shared response as a latent variable over multi-view data. Second, we develop a probabilistic generative model under this framework for modeling and estimating the view specific bases W_i and the shared response latent variable S. A critical aspect of the model is that it directly estimates low dimensional shared features. This is in contrast to methods where the number of features equals the number of voxels [60, 83]. Moreover, the Bayesian nature of the approach provides a natural means of incorporating prior domain knowledge.

Prior Publications Parts of this chapter have been published in [27].

2.2 Shared Response Framework

Here we explore a particular type of multi-view dataset such that data from different views are different realizations of the same underlying source. Our goal is to identify the latent variables that's shared across the views.

We assume that the multi-view data takes the form of a matrix $X_i \in \mathbb{R}^{v_i \times d}$, $i = 1{:}m$. Here m is the number of views, d is the number of time samples, and v_i is the dimension of the observation from the i-th view. Our objective is to model each view as $X_i = W_i S + E_i$ where $W_i \in \mathbb{R}^{v_i \times k}$ is a basis for the i-th view, k is a parameter selected by the experimenter, $S \in \mathbb{R}^{k \times d}$ is a corresponding time series of shared response coordinates, and E_i is an error term, $i = 1{:}m$. To ensure uniqueness of coordinates, it is necessary that W_i has linearly independent columns. We make a stronger assumption that each W_i has orthonormal columns, $W_i^T W_i = I_k$. Properties of this assumption are discussed in §2.2.1.

Two approaches for estimating the bases W_i and the shared response S are illustrated below:

$$\begin{aligned} &\min_{W_i,S} && \textstyle\sum_i \|X_i - W_i S\|_F^2 \\ &\text{s.t.} && W_i^T W_i = I_k, \end{aligned} \tag{2.1}$$

$$\begin{aligned} &\min_{W_i,S} && \textstyle\sum_i \|W_i^T X_i - S\|_F^2 \\ &\text{s.t.} && W_i^T W_i = I_k, \end{aligned} \tag{2.2}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. For $k \leq v$, (2.1) can be solved iteratively by first selecting initial conditions for W_i, $i = 1{:}m$, and optimizing (2.1) with respect to S by setting $S = \frac{1}{m}\sum_i W_i^T X_i$. With S fixed, (2.1) becomes m separate subproblems

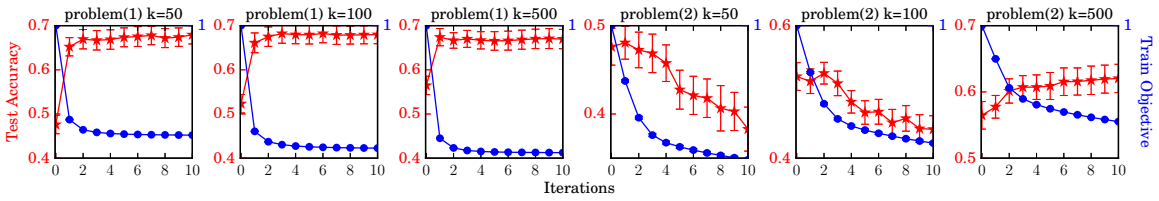


Figure 2.1: Comparison of training objective value and testing accuracy for problem (2.1) and (2.2) over various k on *raider* dataset with 500 voxels of ventral temporal cortex (VT) in image stimulus classificiation experiment (details in §3.4). In all cases, error bars show ± 1 standard error. Figures from [27].

of the form $\min \|X_i - W_i S\|_F^2$ with solution $W_i = \tilde{U}_i \tilde{V}_i^T$, where $\tilde{U}_i \tilde{\Sigma}_i \tilde{V}_i^T$ is an SVD of $X_i S^T$ [61]. These two steps can be iterated until a stopping criterion is satisfied. Similarly, for $k \leq v$, (2.2) can also be solved iteratively. However, for $k < v$, there is no known fast update of W_i given S. Hence this must be done using local gradient decent on the Stiefel manifold [37]. Both approaches yield the same solution when $k = v$, but are not equivalent in the more interesting situation $k \ll v$ (see §2.2.1). What is most important, however, is that problem (2.2) with $k < v$, often learns an uninformative shared response S. This is illustrated in Fig. 2.1 which plots of the value of the training objective and the test accuracy (image classification using the *raider* fMRI dataset, see §3.4)) for a stimulus classification experiment versus iteration count. For problem (2.1), test accuracy increases with decreasing training error, Whereas for problem (2.2), test accuracy decreases with decreasing training error (This can be explained analytically, see §2.2.1). We therefore base our approach on a generalization of problem (2.1). We call the resulting S and $\{W_i\}_{i=1}^m$ a shared response model (SRM).

A simplified view of shared response framework is to view it as a factor model with the ability to include view specific structure. Utilizing the assumption that observations from different views of the same time point are different manifestations of the same shared feature, we concatenate the data over the observed data dimension keeping the temporal dimension synchronized. The factor model perspective learns

a shared latent time series as well as view specific basis. The view specific basis plays the role of bridging between shared feature space and observation space. A simple illustration is shown in Fig. 2.2. This pictorial illustration will help motivate extensions of this framework in §4.

2.2.1 Shared Response Framework Properties

Before extending this simple model, we note a few important properties. First, a solution of (2.1) is not unique. If S, $\{W_i\}_{i=1}^m$ is a solution, then so is QS, $\{W_iQ^T\}_{i=1}^m$, for any $k \times k$ orthogonal matrix Q. This is not a problem as long as we only learn one template and one set of view bases. Any new views or new data will be referenced to the original SRM. However, if we independently learn two SRMs, the group shared responses S_1, S_2, may not be registered (use the same Q). We register S_1 to S_2 by finding a $k \times k$ orthogonal matrix Q to minimize $\|S_2 - QS_1\|_F^2$; then use QS_1 in place of S_1 and W_jQ^T in place of W_j for subjects in the first SRM.

Next, when projected onto the span of its basis, each view's training data X_i has coordinates $S_i = W_i^T X_i$ and the learning phase ensures $S = {}^1/_m \sum_i^m S_i$. The projection to k shared features and the averaging across subjects in *shared feature space* both contribute to across-view denoising during the learning phase. By mapping S back into observation space, we obtain the observation space manifestation W_iS of the denoised, shared component of each view's training data. The training data of the j-th view can also be mapped through the shared response model to the observation space of the i-th view by the mapping $\hat{X}_{i,j} = W_iW_j^T X_j$.

New views are easily added to an existing SRM S, $\{W_i\}_{i=1}^m$. We refer to S as the training template. To introduce a new view $j = m + 1$ with training data X_j, form its orthonormal basis by minimizing the mean squared modeling error $\min_{W_j, W_j^T W_j = I_k} \|X_j - W_jS\|_F^2$. Note that S, and the existing $W_{1:m}$ do not change; we simply add a new view by using its training data under the same stimulus and

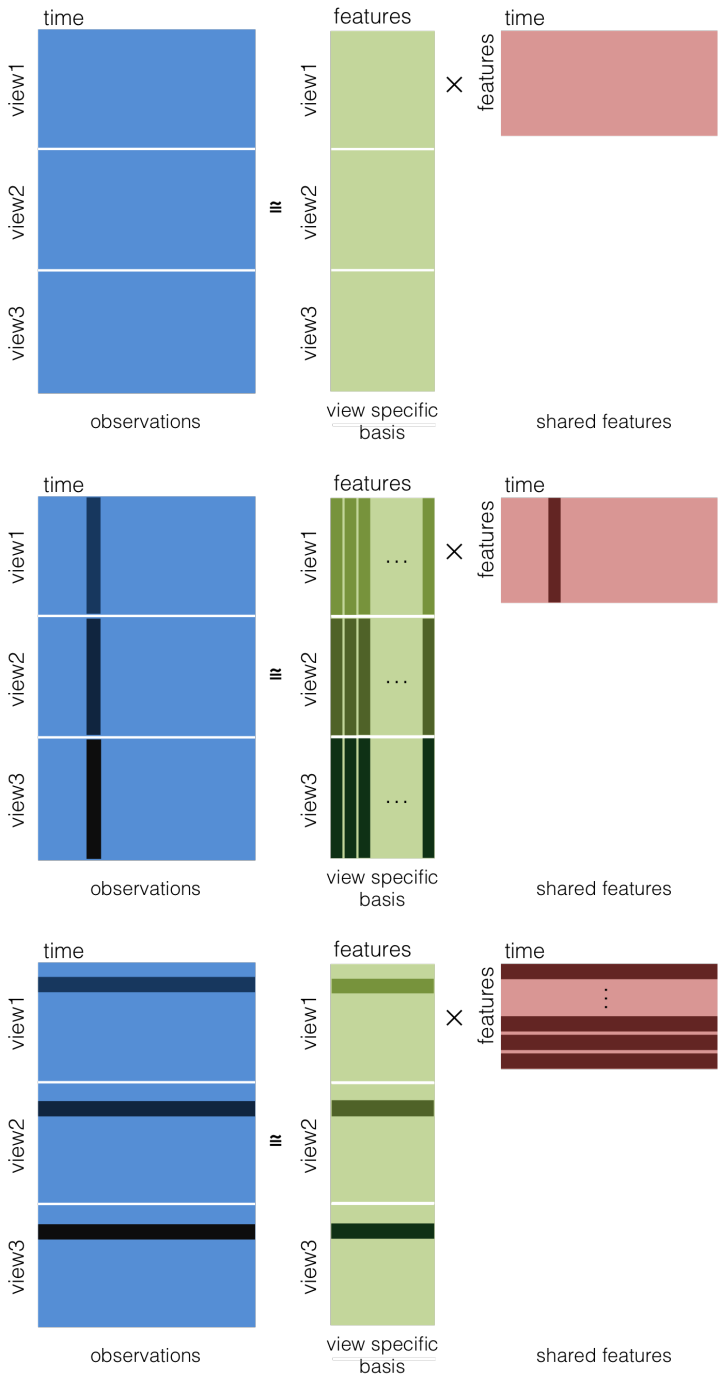


Figure 2.2: Illustration of shared response framework. **Top**: Generic framework. **Middle**: Observations across views of the same timepoint can be viewed as linear combination of view specific basis weighted by a shared feature. **Bottom**: Observed time series of each view can be viewed as linear combination of shared latent time series weighted by subject specific weights.

the template S to determine its view basis. We can also add new data to an SRM. Let X'_i, $i = 1{:}m$, denote new data collected under a distinct stimulus from the same views. This is added to the study by forming $S'_i = W_i^T X'_i$, then averaging these projections to form the shared response for the new data: $S' = {}^1\!/_m \sum_{i=1}^m W_i^T X'_i$. This assumes the learned view specific basis W_i generalizes to the new data. This requires sufficiently rich stimulus in the learning phase.

Comparison between (2.1) **and** (2.2)

When $k = v$ square orthogonal matrices $W_i^T W_i = W_i W_i^T = I$, in this case, we can easily show an identity between (2.1) and (2.2). Starting with the objective function of (2.1), we show that it's equivalent to (2.2):

$$\begin{aligned}
&\|W_i^T X_i - S\|_F^2 \\
=&\mathrm{tr}\big((W_i^T X_i - S)^T (W_i^T X_i - S)\big) \\
=&\mathrm{tr}\big((W_i^T X_i - S)^T W_i^T W_i (W_i^T X_i - S)\big) \\
=&\mathrm{tr}\big((X_i - W_i S)^T (X_i - W_i S)\big) \\
=&\|X_i - W_i S\|_F^2.
\end{aligned} \tag{2.3}$$

However, when $k < v$, W_i is a narrow matrix with orthonormal columns and $W_i W_i^T \neq I$, the equality of (2.3) doesn't hold.

We now consider the difference between (2.1) and (2.2), in an attempt to understand why these objectives can lead to drastically different results (as in Fig. 2.1). X_i can be decomposed as $X_i^W + X_i^{W\perp}$, where $X_i^W = W_i W_i^T X_i$ is the part of X_i in the span of W_i and $X_i^{W\perp} = W_i^\perp W_i^{\perp T} X_i$ is the part of X_i in the orthogonal complement

of span of W_i. By expanding (2.1) and (2.2), we get:

$$\|W_i^T X_i - S\|_F^2 = \text{tr}(X_i^{W^T} X_i^W) - 2\text{tr}(X_i^{W^T} W_i S) + \text{tr}(S^T S) \tag{2.4}$$

$$\|X_i - W_i S\|_F^2 = \text{tr}(X_i^{W^T} X_i^W) + \text{tr}(X_i^{W^\perp T} X_i^{W^\perp}) - 2\text{tr}(X_i^{W^T} W_i S) + \text{tr}(S^T S) \tag{2.5}$$

Since $\text{tr}(X_i^{W^T} X_i^W) + \text{tr}(X_i^{W^\perp T} X_i^{W^\perp}) = \text{tr}(X_i^T X_i)$, (2.5) is trying to find W_i to maximize $\text{tr}(X_i^{W^T} W_i S)$. This maximizes the correlation between transformed observation $W_i^T X_i$ and the shared response S. However, for (2.4), there's a conflict between the first and second terms. The first term is minimizing the variance of projected data, while the second term is maximizing the variance of projected data X_i^W with the shared response. Due to this conflict, (2.2) is prone to find an uninformative basis W_i which doesn't generalize well. This is verified in Fig. 2.1.

From (2.5), the first and second terms add up to $\text{tr}(X_i^T X_i)$, which is a constant value with respoect to W_i. Therefore, we can view maximizing (2.5) as,

$$\max \sum_{i=1}^{m} 2\text{tr}(X_i^T W_i S) - \sum_{i=1}^{m} \text{tr}(S^T S).$$

By replacing $S = {}^1\!/_m \sum_{j=1}^m W_j^T X_j$, we get

$$\begin{aligned} &\frac{1}{m} \sum_{i,j} 2\text{tr}(X_i^T W_i W_j^T X_j) - \frac{1}{m} \sum_{i,j} \text{tr}(X_i^T W_i W_j^T X_j) \\ =&\frac{1}{m} \sum_{i,j} \text{tr}(X_i^T W_i W_j^T X_j). \end{aligned}$$

This shows that the model is maximizing the sum of pair-wise covariance and within view variance after projecting the data into shared feature space through view specific basis.

2.2.2 Discussion of the Orthogonality Constraint

The orthonormal constraint $W_i^T W_i = I_k$ in SRM is similar to that of PCA. In certain special cases, orthogonality can be justified from known neuroscience results. For example, when brain activation patterns for two distinct stimuli classes are known to be spatially disjoint. However, in general, there is no reason to believe that key brain response patterns are orthogonal. So, the orthonormal bases found via SRM should be regarded as a computational tool to aid statistical analysis. From a computational viewpoint, orthogonality has the advantage of robustness and preserving temporal geometry. In §2.2.1, we also show a nice interpretation of SRM's objective function in terms of maximizing within-view variance and paired-wise covariance.

Dropping orthogonality

Here we try dropping the orthogonality constraint in (2.1). This leads to the following modified formulation

$$\min_{W_i,S} \quad \sum_i \|X_i - W_i S\|_F^2. \tag{2.6}$$

(2.6) can be solved iteratively by first selecting initial conditions for W_i, $i = 1{:}m$, and optimizing (2.1) with respect to S by setting $S = (\sum_i W_i^T W_i)^{-1}(\sum_i W_i^T X)$. With S fixed, (2.6) becomes m separate subproblems of the form $\min \|X_i - W_i S\|_F^2$, which is simply ordinary least squares. Hence $W_i = X_i S^T (S S^T)^{-1}$. These two steps can be iterated until a stopping criterion is satisfied.

We empirically evaluate the effect of dropping orthogonality constraint with the experiments in §3.4, including a time segment matching experiment and image category classification. Problem (2.6) is solved for various k. The results are shown in Fig. 2.3. We observe lower predictive performance comparing to the results in Fig. 3.3.

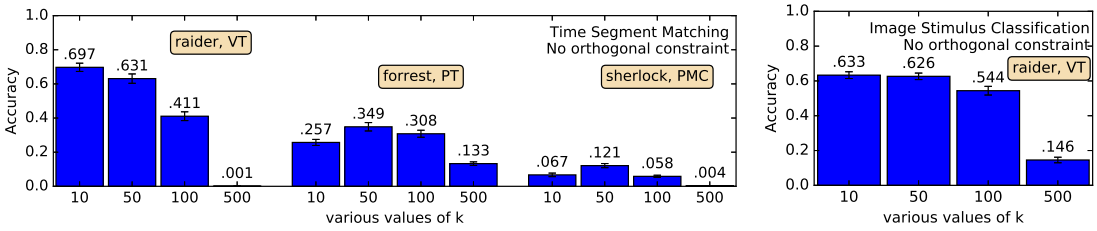


Figure 2.3: Testing accuracy for (2.6) over various k on 500 voxels from *raider*(§3.2.1), *forrest*(§3.2.4), and *sherlock*(§3.2.2) in time segment matching experiment (left) and image stimulus classification experiment (right) (details in 3.4). In all cases, error bars show 1 standard error.

Furthermore, (2.6) tends to perform poorly when k is set to a large value, this might be due to the lack of robustness after dropping the orthogonality constraint.

2.3 Probabilistic Shared Response Model (SRM)

We now extend our simple shared response model to a probabilistic setting. Let $\mathbf{x}_{it} \in \mathbb{R}^v$ denote the observation from the i-th view at time t. For the moment, assume these observations are centered over time. Let $\mathbf{s}_t \in \mathbb{R}^k$ be a hyperparameter modeling the shared response at time $t = 1{:}d$, and model the observation at time t for dataset i as the outcome of a random vector:

$$\mathbf{x}_{it} \sim \mathcal{N}(W_i \mathbf{s}_t, \rho^2 I), \qquad \text{with} \quad W_i^T W_i = I_k, \tag{2.7}$$

where, $\mathbf{x}_{it}$ takes values in $\mathbb{R}^v$, $W_i \in \mathbb{R}^{v \times k}$, $i = 1{:}m$, and ρ^2 is a subject independent hyperparameter. The negative log-likelihood of this model is

$$\mathcal{L} = \sum_t \sum_i \frac{v}{2} \log 2\pi + \frac{v}{2} \log \rho^2 + \frac{\rho^{-2}}{2} (\mathbf{x}_{it} - W_i \mathbf{s}_t)^T (\mathbf{x}_{it} - W_i \mathbf{s}_t).$$

Noting that $\mathbf{x}_{it}$ is the t-th column of X_i, we see that minimizing $\mathcal{L}$ with respect to W_i and $S = [\mathbf{s}_1, \ldots, \mathbf{s}_d]$, requires the solution of:

$$\min \sum_t \sum_i (\mathbf{x}_{it} - W_i \mathbf{s}_t)^T (\mathbf{x}_{it} - W_i \mathbf{s}_t) = \min \sum_i \|X_i - W_i S\|_F^2.$$

Thus maximum likelihood estimation for this model matches (2.1).

In many practical multi-view datasets, $d \gg m$. Since $\mathbf{s}_t$ is time specific but shared across the m views, we see that there is palpable value in regularizing its estimation. In contrast, view specific variables such as W_i are shared across time, a dimension in which data is relatively plentiful. Hence, a natural extension of (2.7) is to make $\mathbf{s}_t$ a shared latent random vector $\mathbf{s}_t \sim \mathcal{N}(\mathbf{0}, \Sigma_s)$ taking values in $\mathbb{R}^k$. The observation for dataset i at time t then has the conditional density $p(\mathbf{x}_{it}|\mathbf{s}_t) = \mathcal{N}(W_i \mathbf{s}_t + \mu_i, \rho_i^2 I)$, where the subject specific mean μ_i allows for a non-zero mean and we assume subject dependent isotropic noise covariance $\rho_i^2 I$. This is an extended multi-subject form of factor analysis, but in factor analysis one normally assumes $\Sigma_s = I$.

To form a joint model, let $\mathbf{x}_t^T = [\mathbf{x}_{1t}{}^T \ldots \mathbf{x}_{mt}{}^T]$, $W^T = [W_1^T \ldots W_m^T]$, $\mu^T = [\mu_1^T \ldots \mu_m^T]$, $\Psi = \mathrm{diag}(\rho_1^2 I, \ldots, \rho_m^2 I)$, $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \Psi)$, and $\Sigma_x = W \Sigma_s W^T + \Psi$. Then

$$\mathbf{x}_t = W \mathbf{s}_t + \mu + \boldsymbol{\epsilon}, \tag{2.8}$$

with $\mathbf{x}_t \sim \mathcal{N}(\mu, \Sigma_x)$ taking values in $\mathbb{R}^{mv}$. For this joint model, we formulate SRM as:

$$\begin{aligned} \mathbf{s}_t &\sim \mathcal{N}(0, \Sigma_s), \\ \mathbf{x}_{it}|\mathbf{s}_t &\sim \mathcal{N}(W_i \mathbf{s}_t + \mu_i, \rho_i^2 I), \\ W_i^T W_i &= I_k, \end{aligned} \tag{2.9}$$

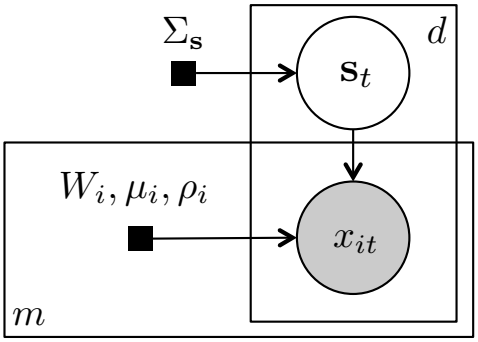


Figure 2.4: Graphical model for SRM. Shaded nodes: observations, unshaded nodes: latent variables, and black squares: hyperparameters. Figure from [27].

where $\mathbf{s}_t$ takes values in $\mathbb{R}^k$, $\mathbf{x}_{it}$ takes values in $\mathbb{R}^v$, and the hyperparameters W_i are matrices in $\mathbb{R}^{v\times k}$, $i = 1{:}m$. The latent variable $\mathbf{s}_t$, with covariance Σ_s, models a shared elicited response across all subjects at time t. By applying the same orthogonal transform to each of the W_i, we can assume, without loss of generality, that Σ_s is diagonal. The SRM graphical model is displayed in Fig. 2.4.

2.4 Parameter Estimation for SRM

There are two ways to estimate the parameters of the SRM model. One way is to apply Stiefel manifold optimization [37] to find maximum likelihood solutions. An alternative is to derive a constrained EM algorithm to find maximum likelihood solutions. This has faster convergence due to coordinate descent. This is the approach followed here. Let θ denote the vector of all parameters. In the E-step, given initial value or estimated value θ^{old} from the previous M-step, we calculate the sufficient statistics by taking an expectation with respect to $p(\mathbf{s}_t|\mathbf{x}_t, \theta^{\text{old}})$:

$$\mathbb{E}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t] = (W\Sigma_s)^T(W\Sigma_s W^T + \Psi)^{-1}(\mathbf{x}_t - \mu), \tag{2.10}$$

$$\begin{aligned}\mathbb{E}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t\mathbf{s}_t^T] &= \text{Var}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t] + \mathbb{E}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t]\mathbb{E}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t]^T \\ &= \Sigma_s - \Sigma_s^T W^T(W\Sigma_s W^T + \Psi)^{-1}W\Sigma_s + \mathbb{E}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t]\mathbb{E}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t]^T. \end{aligned} \tag{2.11}$$

In the M-step, we update the parameter estimate to θ^{new} by maximizing Q with respect to W_i, μ_i, ρ_i^2, $i = 1{:}m$, and Σ_s. This is given by $\theta^{\text{new}} = \arg\max_\theta Q(\theta, \theta^{\text{old}})$, where

$$Q(\theta, \theta^{\text{old}}) = \frac{1}{d}\sum_{t=1}^{d} \int p(s_t|\mathbf{x}_t, \theta^{\text{old}}) \log p(\mathbf{x}_t, s_t|\theta) ds_t.$$

Due to the model structure, Q can be maximized with respect to each parameter separately. To enforce the orthogonality of W_i, we bring a symmetric matrix Λ_i of Lagrange multipliers and add the constraint term $\text{tr}(\Lambda_i(W_i^T W_i - I))$ to the objective function. Setting the derivatives of the modified objective to zero, we obtain the following update equations:

$$\mu_i^{\text{new}} = \frac{1}{d}\sum_t \mathbf{x}_{it}, \tag{2.12}$$

$$W_i^{\text{new}} = A_i(A_i^T A_i)^{-1/2}, \quad A_i = \frac{1}{2}\Big(\sum_t (\mathbf{x}_{it} - \mu_i^{\text{new}})\mathbb{E}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t]^T\Big), \tag{2.13}$$

$$\rho_i^{2\text{new}} = \frac{1}{dv}\sum_t \Big(\|\mathbf{x}_{it} - \mu_i^{\text{new}}\|^2 - 2(\mathbf{x}_{it} - \mu_i^{\text{new}})^T W_i^{\text{new}} \mathbb{E}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t] + \text{tr}(\mathbb{E}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t\mathbf{s}_t^T])\Big), \tag{2.14}$$

$$\Sigma_s^{\text{new}} = \frac{1}{d}\sum_t (\mathbb{E}_{\mathbf{s}|\mathbf{x}}[\mathbf{s}_t\mathbf{s}_t^T]). \tag{2.15}$$

We provide an initial value for W_i by selecting a random matrix in $\mathcal{O}_{v,k}$, $i = 1{:}m$. In our experimental tests (see Chapter 3) we observe robust performance over random initializations.

SRM is adaptively aggregating data with different estimated noise level

We note that (2.1) implicitly assumes subjects had identical noise level. This is reflected by the update equation for S taking a uniform average of the transformed data. In SRM, if instead we set the estimated value of ρ_i^2 to be $\kappa_i^2\lambda^z$ for $0 < \lambda < 1$, $i = 1{:}m$, and let $z \to \infty$, then the shared response becomes a weighted average of the transformed data in which subjects with less noise are weighted more:

$$\lim_{z\to\infty} \mathbb{E}_{\mathbf{s}|\mathbf{x}}[S] = \lim_{z\to\infty} \sum_i \text{diag}\left\{\frac{\kappa_i^{-2}\lambda^{-z}}{\sum_j \kappa_j^{-2}\lambda^{-z} + \sigma_i^{-1}}\right\} W_i^T X_i = \sum_i \frac{1/\kappa_i^2}{\sum_j (1/\kappa_j^2)} W_i^T X_i.$$

2.5 Connections with Related Methods

The shared response framework and SRM are closely connected to various families of factor models with different types of objective function or details in formulation. Those families of factor models include maximizing correlation, maximizing variance/covariance, maximizing independence, structured factor model, and others. In this section, we draw connections among the proposed framework, SRM, and related models. Since there are a large amount of models in each family, we select a few representative ones from each group and conduct a detailed comparison with SRM.

2.5.1 Maximize Correlation (CCA like)

CCA finds two linear transformations for two views such that the correlation between the transformed data is maximized. Probabilistic CCA (pCCA) [13] has been proposed as a probabilistic interpretation of CCA, following a similar approach between probabilistic PCA (pPCA) [117] and PCA. There has also been research using multi-set pCCA by directly extending the results in [13] from 2 sets to multi-set. But to our knowledge, there is no proof connecting multi-set pCCA and multi-set CCA.

A connection between SRM and pCCA can be established in the case of two datasets. In this case, SRM and CCA solutions are different parameterizations of the same pCCA likelihood function, each yielding a maximum of the log-likelihood. We show this result below.

Connections between SRM and CCA

We show that the SRM and CCA solutions are different parameterizations of the two subject pCCA likelihood function. pCCA was proposed as a probabilistic model and it was shown that its maximum likelihood estimates leads to the identical canonical correlation directions obtained in classical two subject CCA. Probabilistic CCA is

proposed as follows:

$$\begin{aligned}
\mathbf{z} &\sim \mathcal{N}(0, I), \mathbf{z} \in \mathbb{R}^k, k \leq v \\
\mathbf{x}_1|\mathbf{z} &\sim \mathcal{N}(W_1\mathbf{z} + \mu_1, \Psi_1) \\
\mathbf{x}_2|\mathbf{z} &\sim \mathcal{N}(W_2\mathbf{z} + \mu_2, \Psi_2),
\end{aligned}$$

where $\mathbf{x}_1$ and $\mathbf{x}_2$ take values in $\mathbb{R}^v$. The maximum likelihood estimates are:

$$\begin{aligned}
\widehat{W_1} &= \widetilde{\Sigma}_{11}\widehat{U}_1\widehat{M_1} \\
\widehat{W_2} &= \widetilde{\Sigma}_{22}\widehat{U}_2\widehat{M_2} \\
\widehat{\Psi}_1 &= \widetilde{\Sigma}_{11} - \widehat{W_1}\widehat{W_1}^T \\
\widehat{\Psi}_2 &= \widetilde{\Sigma}_{22} - \widehat{W_2}\widehat{W_2}^T \\
\hat{\mu}_1 &= \tilde{u}_1 \\
\hat{\mu}_2 &= \tilde{u}_2
\end{aligned}$$

where $\widehat{U}_i = \widetilde{\Sigma}_{ii}^{-\frac{1}{2}}\widehat{V}_i$, $\widehat{V}_1\widehat{P}\widehat{V}_2^T$ is an SVD of $\widetilde{\Sigma}_{11}^{-\frac{1}{2}}\widetilde{\Sigma}_{12}\widetilde{\Sigma}_{22}^{-\frac{1}{2}}$, and $\widehat{M_1}, \widehat{M_2}$ are arbitrary matrices such that $\widehat{P} = \widehat{M_1}\widehat{M_2}^T$. $\widehat{U}_i$ is the transformation matrix for dataset i in CCA solutions. The corresponding log-likelihood value is

$$\mathcal{L}|_{\widehat{W},\widehat{\Psi},\widehat{\mu}} = -\frac{2vd}{2}\log 2\pi e - \frac{d}{2}\log|\widetilde{\Sigma}_x|$$

Next, following similar approach as in [13] we can show that a different mode of maximum likelihood estimates of pCCA leads to close relation with SRM. We derive this mode by taking the derivative of the log-likelihood but using a different

parameterization than pCCA. The maximum likelihood estimates are:

$$\begin{aligned}
\overline{W}_1 &= \overline{U}_1\overline{M}_1 \\
\overline{W}_2 &= \overline{U}_2\overline{M}_2 \\
\overline{\Psi}_1 &= \widetilde{\Sigma}_{11} - \overline{W}_1\overline{W}_1^T \\
\overline{\Psi}_2 &= \widetilde{\Sigma}_{22} - \overline{W}_2\overline{W}_2^T \\
\overline{\mu}_1 &= \tilde{u}_1 \\
\overline{\mu}_2 &= \tilde{u}_2
\end{aligned}$$

where $\overline{U}_1\overline{P}\ \overline{U}_2^T$ is an SVD of $\widetilde{\Sigma}_{12}$, and $\overline{M}_1, \overline{M}_2$ are arbitrary matrices such that $\overline{P} = \overline{M}_1\overline{M}_2^T$. $\overline{U}_i$ is the orthogonal transformation matrix for dataset i in SRM solutions. The corresponding log-likelihood value is

$$\mathcal{L}|_{\overline{W},\overline{\Psi},\overline{\mu}} = -\frac{2vd}{2}\log 2\pi e - \frac{d}{2}\log|\widetilde{\Sigma}_x|$$

which is equal to the log-likelihood derived in pCCA. This shows that for two subjects the SRM and CCA solutions are different parameterizations of the same pCCA likelihood.

□

Connections between SRM and Hyperalignment

Hyperalignment (HA) [60], learns a shared representation by rotating subjects' time series responses to maximize inter-subject time series correlation [58]. This has been proved in [83] and can be shown through connections between HA and regularized CCA [134]. The formulation in [60] is based on problem (2.2) with $k = v$ and W_i a $v \times v$ orthogonal matrix. So this method does not directly reduce the dimension of the feature space, nor does it directly extend to this case (see Fig. 2.1). Although

dimensionality reduction can be done posthoc using PCA, [60] shows that this doesn't lead to performance improvement. In contrast, we show in §2.2.1 that selecting $k \ll v$ can improve the performance of SRM beyond that attained by HA [60] when cast into a probabilistic framework.

We show that Hyperalignment [60] is equivalent to (2.2) when $k = v$. Following is the formulation of Hyperalignment. Note that $X_i \in \mathbb{R}^{v \times d}$ here is the transpose of the notation used in [60].

$$\begin{aligned} \min_{R_i} \quad & \textstyle\sum_{i<j} \|X_i^T R_i - X_j^T R_j\|_F^2 \\ \text{s.t.} \quad & R_i^T R_i = R_i R_i^T = I_v, \end{aligned} \tag{2.16}$$

Let $G = {}^1\!/_m \sum_i X_i^T R_i$, from the equality $\sum_{i<j} \|X_i^T R_i - X_j^T R_j\|_F^2 = \sum_i \|X_i^T R_i - G\|_F^2$ [61], by letting $W_i = R_i$ and $S = G^T$, we get identical formulation as (2.2) when $k = v$.

□

Connections between SRM and regularized Hyperalignment

We now show the difference between SRM and regularized HA (rHA) in [134]. rHA makes a connection between HA and CCA [63] using a ridge CCA formulation [124]. We show rHA on the left and a matching formulation in SRM notation on the right:

$$\begin{aligned} \min \quad & \textstyle\sum_{i<j} \|X_i^T R_i - X_j^T R_j\|_F^2 \\ \text{s.t.} \quad & R_i^T((1-\alpha) X_i X_i^T + \alpha I) R_i = I \end{aligned} \quad \equiv \quad \begin{aligned} \min \quad & \textstyle\sum_i \|W_i^T X_i - S\|_F^2 \\ \text{s.t.} \quad & W_i^T((1-\alpha) X_i X_i^T + \alpha I) W_i = I. \end{aligned}$$

rHA introduces a parameter α bridging the HA constraint $R_i^T R_i = I$ and the CCA constraint $R_i^T X_i^T X_i R_i = I$. rHA becomes standard HA when $\alpha \to 1$ and CCA when

$\alpha \to 0$. In contrast to the regularization on the loading matrices W_i imposed by rHA, SRM introduces regularization on the shared randomness $\mathbf{s}_t$.

□

2.5.2 Maximize Variance/Covariance (PCA like)

For one subject, (2.1) is identical to PCA, and SRM is similar to a variant of pPCA [2] that imposes an orthogonality constraint on the loading matrix. pPCA yields an orthogonal loading matrix. However, due to the increase in model complexity to handle multiple datasets, SRM has an explicit constraint of orthogonal loading matrices. Furthermore, in a standard PCA setting, there's no notion of multi-view data. Therefore, PCA treats all data from a single view perspective and learns principal directions to maximize variance after projection. On the contrary, in essence, SRM takes the multi-view aspect of data into account and learns view specific basis to maximize the sum of both within-view variance and pair-wise covariance (§2.2.1).

2.5.3 Maximize Independence (ICA like)

The GICA, IVA algorithms [91] do not assume time-synchronized stimulus and hence concatenate data along the time dimension (implying spatial consistency) and learn spatial independent components. In contrast, we assume a time-synchronized stimulus for anchoring the shared response to overcome a spatial mismatch in view specific bases.

Connections between SRM and Standard ICA

Standard ICA [67] is a factor model that tries to find a linear representation of data so that the components are statistically independent, or as independent as possible.

Using our notation, given data $\mathbf{x}$,

$$\mathbf{x} = W\mathbf{s},$$

where $\mathbf{s}$ is the independent components. There are two main differences between ICA and SRM. First, ICA isn't designed for multiple datasets. Although there are multiple datasets extension of ICA, such as GICA, IVA. Second, ICA doesn't have the notion of "shared response". It's maximizing statistical independence, but this doesn't necessarily lead to shared components. This is examined further in Chapter 3.5.

□

2.5.4 Structured Factor Model

Topographic Factor Analysis (TFA) [85] is a factor model using a topographic basis composed of spherical Gaussians with different centers and widths. This choice of basis is constraining, but since each factor is a "Gaussian blob" in the brain, it has the advantage of providing a simple spatial interpretation. Hierarchical TFA [84] extends [85] into group setting such that each subject has a set of topographic basis, and subject level topographic bases are a perturbation of a group level template topographic basis. These models can be trained with either variational inference or maximum a posteriori (MAP) inference.

On the other end, instead of parameterizing the model with a structured basis, [1, 121, 38, 36] propose a series of dictionary-learning-based models with structural sparsity regularization. The models regularize the basis to be piece-wise smooth and compact, making up blobs, contrary to the scattered activation patterns.

We have tried extending SRM in both directions, model parameterization for a structural basis and regularization for a structural basis. However, learning a basis

that is both structured and useful for prediction purpose turns out to be very difficult. We obtain results which are either a structural basis with low predictive performance or nonstructural basis with a high predictive performance. Bases with high predictive power, but nonstructural, are useful for detecting the existence of information within ROI. However, structured bases with low predictive power aren't useful, since we can't verify that the structure is informative. How one can strike a nice balance or establishing some form of trade-off between structural basis and predictive power will be an interesting future direction.

2.5.5 Other Methods

SRM is also related to ridge regression. We make this connection by showing that single subject SRM is connected with ridge regression with an orthogonality constraint on the loading matrix.

Connections between SRM and Ridge Regression

SRM is related to ridge regression. We make this connection Assume $\mathbf{s}_t$ is sampled from $\mathcal{N}(0, \gamma^2 I)$ with γ^2 known, and that $\Sigma_{x'_m} = I$. When $M = 1$, MAP estimation of W_i and $\mathbf{s}_t$, $t = 1{:}T$, estimates a mode of the log posterior distribution $\sum_t \log p(\mathbf{s}_t|\mathbf{x}_{it})$:

$$\max \sum_t \big(\log p(\mathbf{x}_{it}|\mathbf{s}_t) + \log p(\mathbf{s}_t) \big) \equiv \min \sum_t \big(\|\mathbf{x}_{it} - W_i \mathbf{s}_t\|_F^2 + \gamma^{-2} \|\mathbf{s}_t\|_2^2 \big).$$

This is ridge regression for $\mathbf{s}_t$ given W_i, and least squares regression for W_i (with an orthogonality constraint), given $\mathbf{s}_t$, $t = 1:T$. In the multi-subject case, MAP estimation of W_i and $\mathbf{s}_t$, will be similar but with a block-wise orthonormal structure in W:

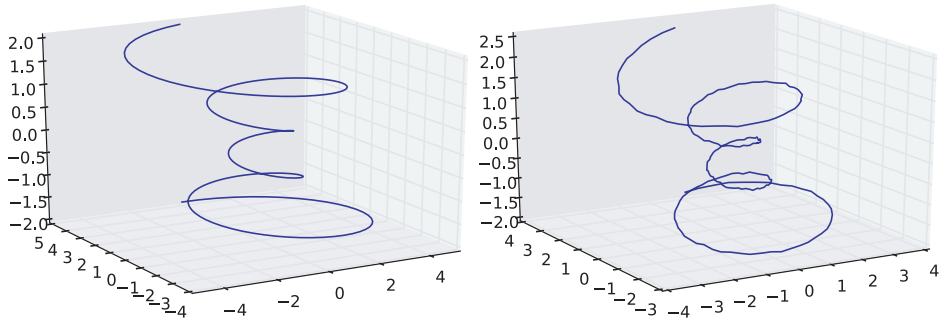
$$\max \sum_t \sum_i \big(\log p(\mathbf{x}_{it}|\mathbf{s}_t) + \log p(\mathbf{s}_t) \big) \equiv \min \sum_t \big(\|\mathbf{x}_t - W \mathbf{s}_t\|_F^2 + \gamma^{-2} \|\mathbf{s}_t\|_2^2 \big).$$

□

2.6 Empirical Test of SRM with Synthetic Data

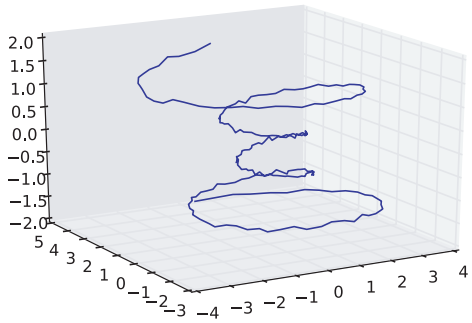
We begin the evaluation of SRM's ability to reconstruct a shared response by using a synthetic dataset. We first generate the three-dimensional shared response shown in Fig. 2.5(i) with 200 time points. Three-dimensional shared response is used due to its easy visualization. We then synthesize observations for five different views from this three-dimensional shared response. For each view, we independently augment the three-dimensional shared response with 30 independently sampled noise features. The noise features play the role of unshared view specific responses. We then rotate each view's dataset using a view specific random orthogonal matrix. So for each view, we obtain a data matrix $X_i \in \mathbb{R}^{33\times 200}$ that contains a mixture of the three-dimensional shared response and 30 channels of independent noise. Finally, SRM is fit to this synthetic data with various values of the signal-to-noise ratio. The SNR in dB is defined as $10\log(P_{\text{signal}}/P_{\text{noise}})$. According to the preset SNR and the signal power, we calculate the corresponding noise power using the above equation. The noise is generated from a zero mean Gaussian with variance as the square root of calculated noise power. In this experiment, k is directly set to the known ground truth, three dimensions. For real datasets, cross-validation will be used to select k.

Due to the non-identifiability property of the model (§2.2.1), the shared response is only estimated up to an orthogonal transformation. We use the approach described in §2.2.1 to learn an extra rotational matrix Q to find the rotation that best matches the learned shared response with the ground truth . This is necessary for visually comparing the reconstructed shared response and the ground truth shared response. The final results (Fig. 2.5) show that the method is capable of reconstructing the original shared response up to a rotation.

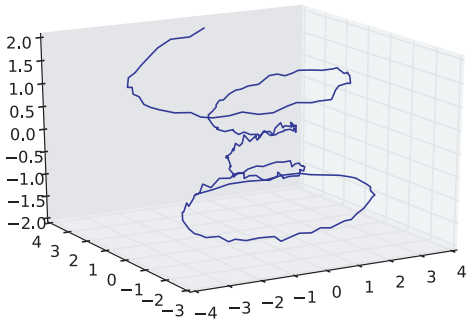


(i) True shared response.

(ii) 10dB SNR



(iii) 5dB SNR



(iv) 1dB SNR

Figure 2.5: Estimated shared features from the synthetic dataset. (i) true shared response. Estimate of shared response under (ii) 10dB SNR; (iii) 5dB SNR; and (iv) 1dB SNR.

2.7 Discussion and Conclusion

In this chapter, we have proposed a general multi-view learning framework for identifying shared response as a latent time series over multi-view data. This framework can be viewed as a multi-view factor model, such that it decomposes data from each view into sets of view specific basis as well as a set of shared response that is shared across views. This framework will guide our thinking for several models in this thesis. The low-dimensional shared representation also helps with both denoising and learning a concise representation.

We utilize the assumption that data from different views of the same timepoints are different realizations of the same underlying source to accommodate non-uniform dimensionality across views. However, this might not be the case for some datasets. We demonstrate an approach to overcome this requirement in §3.9.1.

Besides the general framework, we also develop a probabilistic generative model, shared response model (SRM), for modeling and estimating the view specific basis and the shared response latent variable such that the within-view variance and across views pair-wise covariance is maximized. The idea of a probabilistic generative model roots from the two different approaches for estimating the bases in the initial framework. In that framework, we demonstrate the superior predictive performance of the generative approach. To do inference with SRM, a constrained EM algorithm is developed preserving the orthogonality constraint of view specific basis during inference.

Finally, connections between SRM and related methods are drawn in §2.5 and in §2.6 an experiment on synthetic data is conducted, demonstrating the models' effectiveness in reconstructing the synthetic shared response.

Chapter 3

Shared Response Model on Neuroimaging Data

3.1 Introduction

In Chapter 2, we motivated SRM from a multi-view learning perspective without tying it to a specific application. In this chapter, we explore its application in neuroimaging. In addition to studying the application of SRM to fMRI data, this chapter is a thorough reference guide for neuroscientists on using SRM with neuroimaging data.

To conduct neuroimaging analysis with SRM, we start with a series of questions that the *users* of SRM should ask before using it.

What are the views? From the multi-view learning perspective, a view can be of various forms. For example, we can treat neuroimaging data from different subjects, different modalities, different areas of the same subject, or even stimulus features as different "views." This question is highly related to the input data we want to analyze. As an example, in fMRI studies, it's typical to use a multi-subject dataset. This allows verification of generalizability of the scientific discovery, and increases statisti-

cal power. In this case, to use a multi-subject dataset in the multi-view representation learning framework, we treat each subject's fMRI data as a different view.

What is the hypothesis that we are testing? A fundamental approach to scientific research is the idea of testing a hypothesis. Depending on the hypothesis, there are different ways to use SRM. Here we explain a perspective we hold on using prediction as hypothesis testing. We argue that some forms of prediction can be viewed as hypothesis testing. For example, if we want to test whether image category information exists in ventral temporal cortex (VT), we could conduct image category prediction using the fMRI response from ventral temporal cortex. The null hypothesis will be that image category information doesn't exist in VT and the prediction accuracy of image category should be at chance level. For every batch of data, we conduct classification and a classification accuracy is calculated. With the classification accuracy and the chance level accuracy, a binomial test can be conducted. Through different experiments in this chapter, we show different ways to use SRM.

Which space are we analyzing in? For SRM, there is one voxel space per view and a shared feature space. A critical decision when using SRM is in which space to analyze the data. Shared feature space and voxel space exhibit very different properties. Shared feature space tends to be lower dimensional, and the low-dimensional representation generally leads to higher predictive performance. However, the abstract shared feature space doesn't have the notion of "voxel," which makes associating predictive performance with brain regions hard. On the other hand, voxel space has exactly the same dimensionality as the number of voxels. The key advantage of analyzing in the voxel space is its close association with brain anatomy, which provides high interpretability. Empirically, conducting statistical analysis in voxel space tends to have lower sensitivity.

With these questions answered, we can then proceed to how we use SRM on fMRI data. To conduct multivariate statistical analysis one might select a subset of voxels within an anatomical region of interest (ROI), or select a subset of principal components of the ROI, then use these selected features for subsequent analysis. In a similar way, the fMRI data aggregation problem, can be thought of as a two step process. First use training data to learn a mapping of each subject's measured data to a shared feature space in a way that captures the across-subject shared response. One can then use these learned mappings to project held out data for each subject into the shared feature space and perform a statistical analysis. Alternatively one can use the learned mappings to project data from the feature space to voxel space for analysis. We can view analysis in the feature space as working with a denoised representation of the data. Empirically, working in the feature space leads to higher predictive power. However, the drawback of working in the feature space is that we are disconnected from the anatomical information which can be critical for neuroscience interpretability. There are two ways to solve this. One way is to use a searchlight [77, 41, 50, 139, 29] approach. This is demonstrated in §3.7. The other way is to conduct the analysis in voxel space by first mapping to feature space and then mapping back to voxel space. This is demonstrated in §3.6.

To make this more precise, let multi-subject fMRI time-series data $X_i \in \mathbb{R}^{v_i \times d}$, $i = 1{:}m$, be collected for m subjects as they are presented with identical, time synchronized stimuli. Here d is the number of time samples in TRs (Time of Repetition), and v_i is the number of voxels. Note that the number of voxels across subjects/views does not have to be the same, allowing greater flexibility for data analysis. The model learns view specific bases that bridge between view observation spaces and a shared feature space.

In this chapter, we demonstrate various ways to apply SRM for fMRI data analysis. SRM learns a low dimensional shared feature space such that between views pair-wise

covariance and within view variance is maximized after projecting to the feature space. The key "product" is the mapping from each view's observation space to the shared feature space. As long as we hypothesize there's some latent representation shared across different views, we can apply SRM and test the hypothesis through predictive experiments. With this in mind, we first show how to use SRM for identifying shared feature space across subjects. We show generalization to a new subject and a new stimulus through comparison with spatial smoothing, using time segment matching, and image classification. In addition to identifying what is shared across subjects, we can also remove the shared response from the original response to obtain the residual individual response. This is demonstrated in the differentiating between groups experiment in §3.5.

Prior Publications and Acknowledgment Parts of this chapter have been published in [27, 139]. I thank Michael J. Arcaro for his permission to use retinotopy figures in §3.6.

3.2 fMRI Datasets

We assess the performance and robustness of SRM using the fMRI datasets shown in Table 3.1. These were collected using different MRI machines, subjects, and preprocessing pipelines. For the various experiments, we either use data from whole brain, a region of interest (ROI), or in searchlights. The areas that are being used will be specify in the experiment description.

3.2.1 Raider dataset

The *raider* [60] dataset has two parts. The first part was collected while 10 subjects viewed the movie "Raiders of the Lost Ark" (110 mins), and the second part was

collected while the same 10 subjects viewed a series of still images (7 categories, 8 runs).

3.2.2 Sherlock dataset

The *sherlock* [25] dataset has two parts. The first part, *sherlock-movie*, is a movie watching part, the dataset was collected while 16 subjects watched an episode of the BBC TV series "Sherlock" (50 mins). The second part, *sherlock-recall* is a movie free-recall experiment. This was collected while the same 16 subjects verbally reiterated the "Sherlock" episode without any outside prompts.

3.2.3 Audiobook dataset

The *audiobook* [138] dataset was collected while 40 subjects listened to a narrated story (15 mins) with two possible interpretations. Half of the subjects had a prior context favoring one interpretation, the other half had a prior context favoring the other interpretation. Post scanning questionnaires showed no difference in comprehension but a significant difference in interpretations between groups.

3.2.4 Forrest dataset

The *forrest* [52] dataset was collected while 18 subjects listened to an auditory version of the film "Forrest Gump" (120 mins).

3.2.5 Region of interest (ROI)

We use different ROI for different experiment. The ROIs vary in function from visual processing, language, memory, to understanding others' mental states. The regions that we use include visual cortex, ventral temporal cortex (VT) [59], planum temporale (PT) [48], posterior medial cortex (PMC) [87], and default mode network

Dataset	Subjs	TRs (s/TR)
sherlock (audio-visual movie) [25]	16	1976 (1.5)
raider (audio-visual movie) [60]	10	2203 (3)
forrest (audio movie) [52]	18	3599 (2)
audiobook (narrated story) [138]	40	449 (1.5)

Table 3.1: fMRI datasets and their respective properties.

(DMN) [105]. With respect to the different datasets, the number of voxels for each ROI ranges from several hundred voxels to few thousand voxels.

3.3 SRM and Spatial Smoothing

In this experiment, we compare two approaches, SRM and spatial smoothing, to detect a shared response between two independent groups receiving the same stimulus. From a multi-view perspective, each subject is a view and we compare the similarity of estimated within-group shared response.

We first use spatial smoothing to determine if we can detect a shared response in posterior medial cortex (PMC) [87] for the *sherlock* dataset (§3.2.2). The subjects are randomly partitioned into two equal sized groups, the data for each group is averaged, we calculate the Pearson correlation over voxels between these averaged responses for each time, then average these correlations over time. This is a measure of similarity of the sequence of brain maps in the two average responses. We repeat this for five random subject divisions and average the results. If there is a shared response, we expect a positive average correlation between the groups, but if functional topographies differ significantly across subjects, this correlation may be small. If the result not distinct from zero, a shared response is not detected. The computation yields the benchmark value 0.26 ± 0.006 shown as the purple bar in the right plot in Fig. 3.1. This is support for a shared response in PMC, but we posit that the subject's functional topographies in PMC are misaligned. To test this, we use a Gaussian filter,

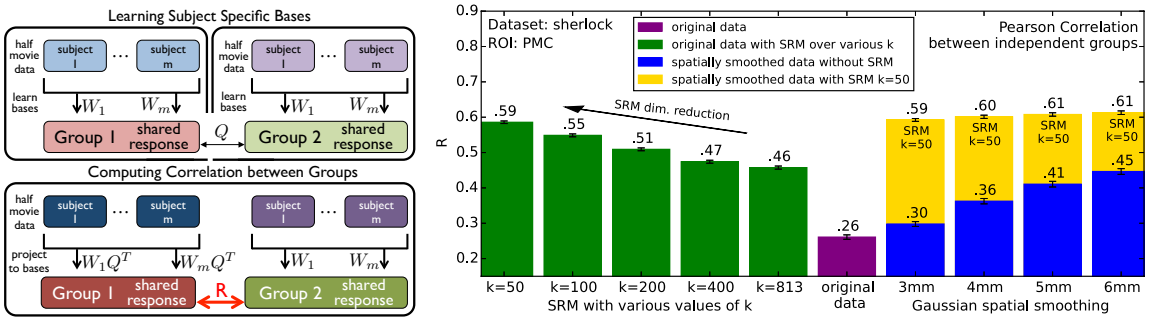


Figure 3.1: SRM and spatial smoothing. Left: Learn using half of the data, then compute between group correlation on other half. Right: Pearson correlation after spatial smoothing, and SRM with various k. Error bars: ± 1 stand. error. Figures from [27].

with width at half height of 3, 4, 5 and 6mm, to spatially smooth each subject's fMRI data. We then recalculate the average Pearson correlation as described above. The results, shown as blue bars in Fig. 3.1, indicate higher correlations with greater spatial smoothing. This indicates greater average correlation of the responses at lower spatial frequencies, suggesting a fine scale mismatch of functional topographies across subjects.

We now test the robustness of SRM using the unsmoothed data. The subjects are randomly partitioned into two equal sized groups. The data in each group is divided in time into two halves, and the same half in each group is used to learn a shared response model for the group. The independently obtained group templates S_1, S_2, are then registered using a $k \times k$ orthogonal matrix Q (method outlined in §2.2.1: finding a $k \times k$ orthogonal matrix Q to minimize $\|S_2 - QS_1\|_F^2$; then use QS_1 in place of S_1 and W_jQ^T in place of W_j for subjects in group 1. This leaves the product W_jS_1 invariant for subjects in that group.) For each group, the second half of the data is projected to feature space using the subject-specific bases and averaged. Then the Pearson correlation over features is calculated between the group averaged shared responses, and averaged over time. This is repeated using the other halves of the subject's data for training and the results are averaged. The average results over 5

random subject divisions are report as the green bars in Fig. 3.1. With $k = 813$ there is no reduction of dimension and SRM achieves a correlation equivalent to 6mm spatial smoothing. This strong average correlation between groups, suggests some form of shared response. As expected, if the dimension of the feature space k is reduced, the correlation increases. A smaller value of k, forces SRM to focus on shared features yielding the best data representation and gives greater noise rejection. Learning 50 features achieves a 33% higher average correlation in feature space than is achieved by 6mm spatial smoothing in voxel space. A commensurate improvement occurs when SRM is applied to the spatially smoothed data. We note that higher correlation doesn't necessarily reflect the usefulness of learned basis in prediction. Ideally we will apply SRM on unsmoothed data for estimation of subject specific basis, which can be viewed as an adaptive smoothing. Smoothed data is not ideal due to the removal of fine-grained spatial patterns. We also expect a higher predictive performance by using unsmoothed data compared to smoothed data because the former contains more information that can be utilized by SRM.

3.4 Temporal Similarity and Image Classification

In this set of experiments, we test if the shared response estimated by SRM generalizes to new subjects and new data using versions of two experiments from [60] (unlike in [60], here the held out subject is not included in learning phase). Different subjects are treated as different views in this experiment.

The first experiment tests if an 18s time segment from a held-out subject's new data can be located in the corresponding new data of the training subjects. A shared response and subject specific bases are learned using half of the data, and the held out subject's basis is estimated using the shared response as a template. Then a random 18s test segment from the unused half of the held out subject's data is projected onto

the subject's basis. We then locate the 18s segment in the averaged shared response of the other subject's new data that is maximally correlated with the test segment (see Fig. 3.2). The held out subject's test segment is correctly located (matched) if its correlation with the average shared response at the same time point is the highest; segments overlapping with the test segment are excluded. We record the average accuracy and standard error by two-fold cross-validation over the data halves and leave-one-out over subjects.

The results using three different fMRI datasets, *raider* (§3.2.1), *forrest* (§3.2.4), and *sherlock* (§3.2.2), with distinct ROIs are shown in the top plot of Fig. 3.3. The accuracy is compared using: anatomical alignment (MNI [88], Talairach (TAL) [114]); standard PCA, and ICA feature selection (FastICA implementation [67]); the Hyperalignment (HA) method [60]; and SRM. PCA and ICA are directly applied on joint data matrix $X^T = [X_1^T \dots X_m^T]$ for learning W and S, where $X \approx WS$ and $W^T = [W_1^T \dots W_m^T]$. SRM demonstrates the best matching of the estimated shared temporal features of the methods tested. This suggests that the learned shared response is more informative of the shared brain state trajectory at an 18s time scale. Moreover, the experiment verifies generalization of the estimated shared features to subjects not included in the training phase and new (but similar) data collected during the other half of the movie stimulus. Since we expect accuracy to improve as the time segment is lengthened, what is important is the relative accuracy of the compared methods. The method in (2.1) can be viewed as non-probabilistic SRM. In this experiment, it performs worse than SRM but better than the other compared methods. The effect of the number of features used in SRM is shown in Fig. 3.3, lower left. This can be used to select k. A similar test on the number of features used in PCA and ICA indicates lower performance than SRM (results not shown). Both PCA and ICA are used by concatenating subjects' response matrices along the spatial dimension (implying temporal synchrony). Then applying PCA/ICA on the concatenated

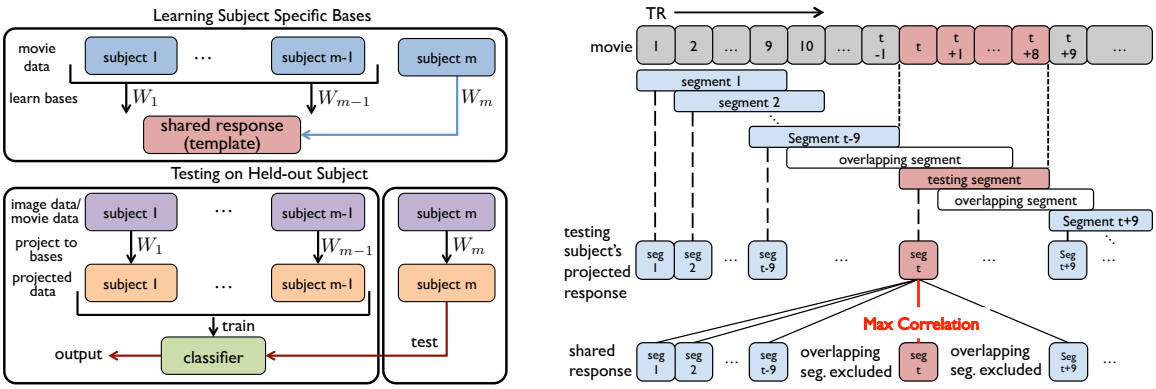


Figure 3.2: Left: Learn subject specific bases. Test on held out subject and data. Right: Time segment matching by correlating with 9 TR segments in the shared response. Figures from [27].

joint data matrix. However, this leads to the unnatural constraint that the learned subject specific basis are jointly orthogonal. Another approach is to concatenate data along the temporal dimension. This has been has been discussed in a hyperalignment setting in [28]. It leads to improved run time but not to an improvement in prediction performance.

We now use the image viewing data and the movie data from the *raider* dataset (§3.2.1) to test the generalizability of a learned shared response to a held-out subject and new data under a very distinct stimulus. The *raider* movie data is used to learn a shared response model, while excluding a held-out subject. The held-out subject's basis is estimated by matching its movie response data to the estimated shared response. The effectiveness of the learned bases is then tested using the image viewing dataset [60]. After projecting the image data using the subject bases to feature space, an SVM classifier is trained and the average classifier accuracy and standard error is recorded by leave-one-out across subject testing. The results, lower right plot in Fig. 3.3, support the effectiveness of SRM in generalizing to a new subject and a distinct new stimulus. Under SRM, the image stimuli can be slightly more accurately identified using other subjects' data for training than using a subject's own data, indicating that the learned shared response is informative of image category.

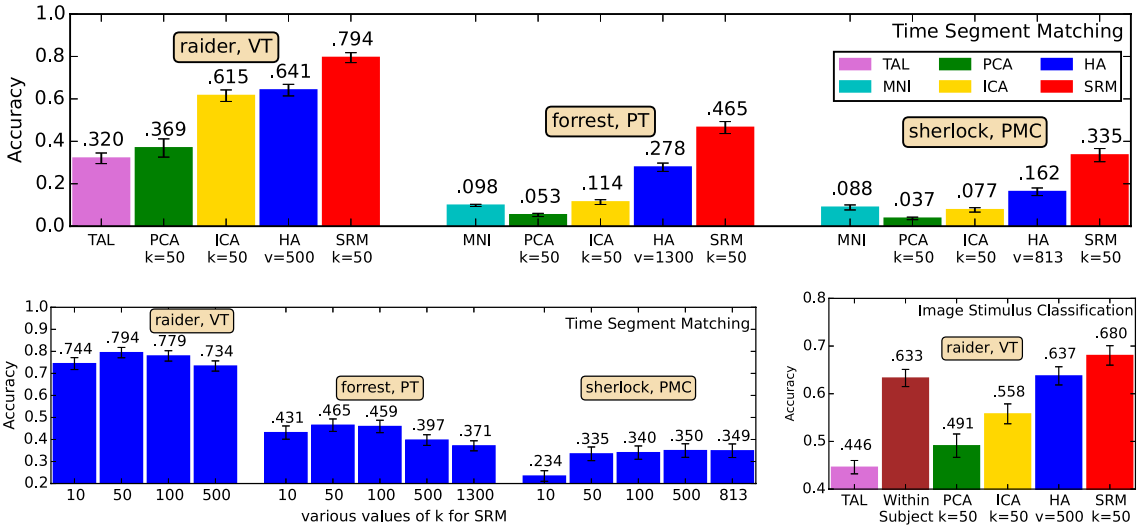


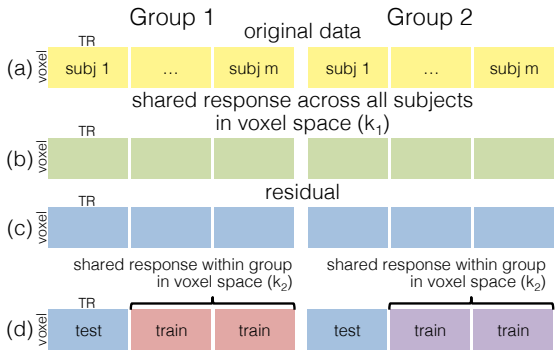
Figure 3.3: Top: Comparison of 18s time segment classification on three datasets using distinct ROIs. Bottom: (Left) SRM time segment classification accuracy vs k. (Right) Learn bases from movie response, classify stimulus category using still image response. For *raider* and *forrest*, we conduct experiment on ROI in each hemisphere separately and then average the results. For *sherlock*, we conduct experiment over whole PMC. The TAL results for the *raider* dataset are from [60]. Error bars: ± 1 stand. error. Figures from [27].

3.5 Differentiating between Groups

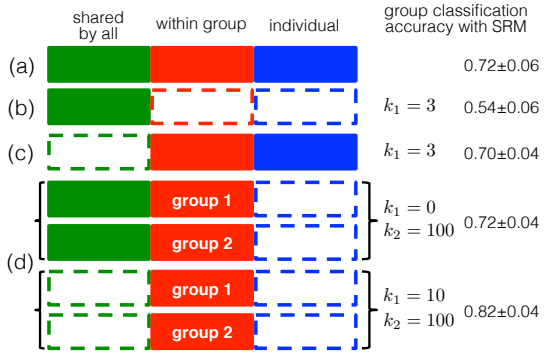
In this experiment, we predict group association of a testing subject from the *audio-book* dataset (§3.2.3). This experiment also demonstrates the potential for the model to decouple shared and individual response leading to better predictive performance. Each subject is treated as a view, and DMN ROIs from all subjects are used.

If subjects are given group labels according to the two prior contexts, a linear SVM classifier trained on labeled voxel space data and tested on the voxel space data of held out subjects, can distinguish the two groups at an above chance level. This is shown as the leftmost bar in the bottom figure of Fig. 3.4. This is consistent with previous similar studies [6].

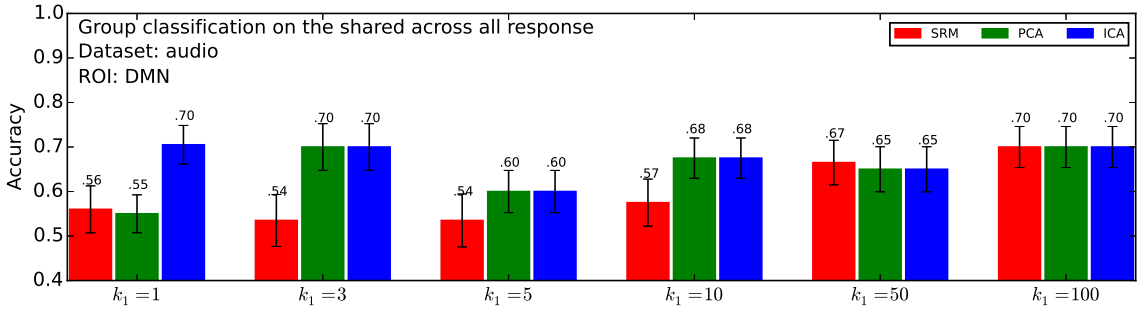
We now test if SRM can distinguish the two subject groups using the procedure outlined in the rows of Fig. 3.4i. We use the original data $X^{\text{g1}}_{1:m}, X^{\text{g2}}_{1:m}$ of all subjects (Fig. 3.4(i)(a)) to learn a k_1-dimensional shared response S^{all} and subject bases



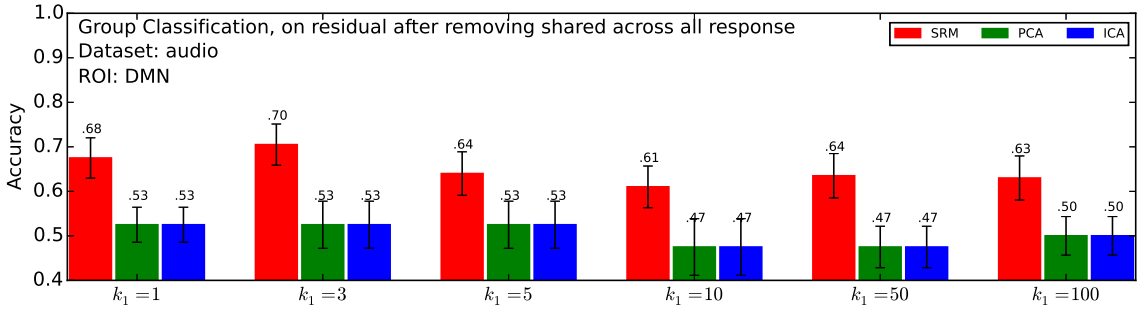
(i) Experiment procedure.



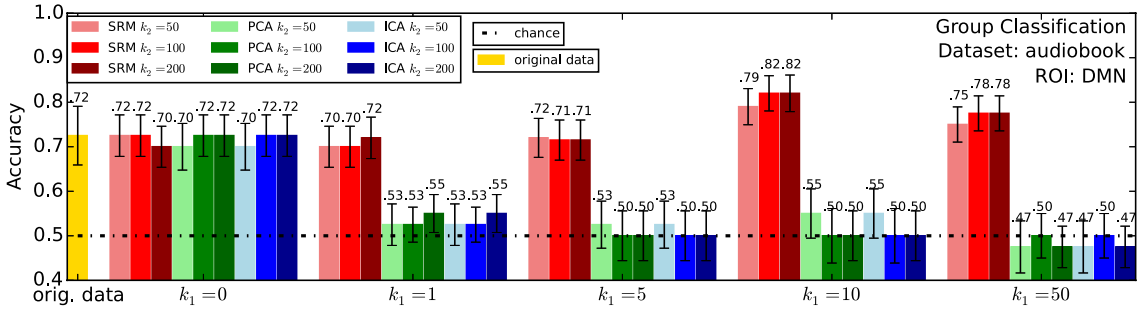
(ii) Cartoon illustration of the experiment.



(iii) On removed shared across all response.



(iv) On residual after removed shared across all response.



(v) After removed shared across all response and train on shared within group response.

Figure 3.4: Procedure, illustration, and results for differentiating between groups experiment. Figures (i,ii,v) from [27].

$W^{\text{all}}_{\text{gj},1:m}$. This shared response is then mapped to voxel space using each subject's learned topography (Fig. 3.4(i)(b)). This is then subtracted from the subject's data to form the residual response $X_i^{\text{gj}} - W^{\text{all}}_{\text{gj},i} S^{\text{all}}$ for subject i in group j (Fig. 3.4(i)(c)). Leaving out one subject from each group, we use two within-group applications of SRM to find k_2-dimensional within-group shared responses $S^{\text{g1}}, S^{\text{g2}}$, and subject bases $W^{\text{g1}}_{1:m}, W^{\text{g2}}_{1:m}$ for the residual response. These are mapped into voxel space $W_i^{\text{gj}} S^{\text{gj}}$ for each subject (Fig. 3.4(i)(d)). The first step application of SRM across all subjects, yields a denoised estimate of the shared response that is used to form the residual response. The subsequent within-group applications of SRM give denoised estimates of the within-group shared response of the residual response. Both applications of SRM seek to remove components of the original response that are uninformative of group membership. Finally, a linear SVM classifier is trained using the voxel space group-labeled data, and tested on the voxel space data of held out subjects. The results are shown as the red bars in Fig. 3.4. When using $k_1 = 10$ and $k_2 = 100$, we observe significant improvement in distinguishing the groups.

One can visualize why this works using the cartoon in Fig. 3.4(ii) showing the data for one subject modeled as the sum of three components: the response shared by all subjects, the response shared by subjects in the same group after the response shared by all subjects is removed, and a final residual term called the individual response (Fig. 3.4(ii)(a)). We first identify the response shared by all subjects (Fig. 3.4(ii)(b)); subtracting this from the subject response gives the residual (Fig. 3.4(ii)(c)). The second within-group application of SRM removes the individual response (Fig. 3.4(ii)(d)). By tuning k_1 in the first application of SRM and tuning k_2 in the second application of SRM, we estimate and remove the uninformative components while keeping the informative component.

Classification using the estimated shared response ($k_1 \leq 10$) results in accuracy around chance (Fig. 3.4(ii)(b)), indicating that it is uninformative for distinguishing

the groups. The classification accuracy using the residual response is statistically equivalent to using the original data (Fig. 3.4(ii)(c)), indicating that only removing the response shared by all subjects is insufficient for improvement. The classification accuracy that results by not removing the shared response ($k_1 = 0$) and only applying within-group SRM (Fig. 3.4(ii)(d)), is also statistically equivalent to using the original data. This indicates that only removing the individual response is also insufficient for improvement. By combining both applications of SRM we remove both the response shared by all subjects and the individual responses, keeping only the responses shared within groups. For $k_1 = 10, k_2 = 100$, this leads to significant improvement in performance (Fig. 3.4(ii)(d) and Fig. 3.4(v)).

For comparison, we performed the same experiment using PCA and ICA in place of SRM (Fig. 3.4). In this case, after removing the estimated shared response ($k_1 \geq 1$) group identification quickly drops to chance since the shared response is informative of group difference (around 70% accuracy for distinguishing the groups (Fig. 3.4(iii), Fig. 3.4(v))). So PCA and ICA can't be relied upon to identify a shared response.

Finally, we provide a comparison of all methods at different steps in the experiment pipeline. The left most yellow bar in Fig. 3.4(v) shows the classification accuracy directly on original data. The above chance results indicates that the groups are distinguishable in DMN ROI.

In Fig. 3.4(iii), we test how informative is the shared across all response in differentiating groups. For SRM, we observe chance level accuracy with $k_1 \leq 10$, but above chance accuracy with $k_1 \geq 50$. The shared by all subjects response is expected to be uninformative for distinguishing between groups. However, when we use large k_1, SRM starts to incorporate shared within group only response as shared by all subjects response, because we force it to identify a large subspace. This is demonstrated by the above chance performance with $k_1 = 100$. On the other hand, for PCA and ICA, we observe above chance accuracy when $k_1 \geq 1$. PCA and ICA identify

components that lead to maximum variance and statistical independence. However, these objective functions do not fit the multi-view learning framework, they do not contain the concept of shareness. So PCA and ICA can't be relied upon to identify a shared response. The above chance accuracies of PCA and ICA suggest that the estimated response are not truly shared by all subjects, which should be uninformative in differentiating between groups.

Fig. 3.4(iv) shows that the residual after removing shared across all response (Fig. 3.4(i)(c)) is uninformative for PCA and ICA. This is because the informative part of original data has been removed, suggesting the incorrect estimation of shared across all response (Fig. 3.4(i)(b)). For SRM, we observe consistent above chance performance in distinguishing groups. The performance is similar to using the whole data. This suggests that the shared across all response estimated by SRM does not contain information that helps differentiating groups.

In Fig. 3.4(v), without removing shared response ($k_1 = 0$), we observe that all three methods are effective in distinguishing between two groups. However, this doesn't lead to better performance than on original data. This suggests removing individual response is insufficient for improvement. With proper selection of k_1 and k_2, we observe statistically significant improvement with $k_1 = 10$ and $k_2 = 100$. This shows that by removing both the shared-by-all subjects response and the individual response, the denoised data demonstrates better distinguishability between the groups.

3.6 Retinotopy Data

In order to have a better understanding of how SRM works, we explore using SRM on a specific type of fMRI data for which we have a good understanding of the ground truth. Specifically, we use polar angle retinotopy data. Polar angle data is collected

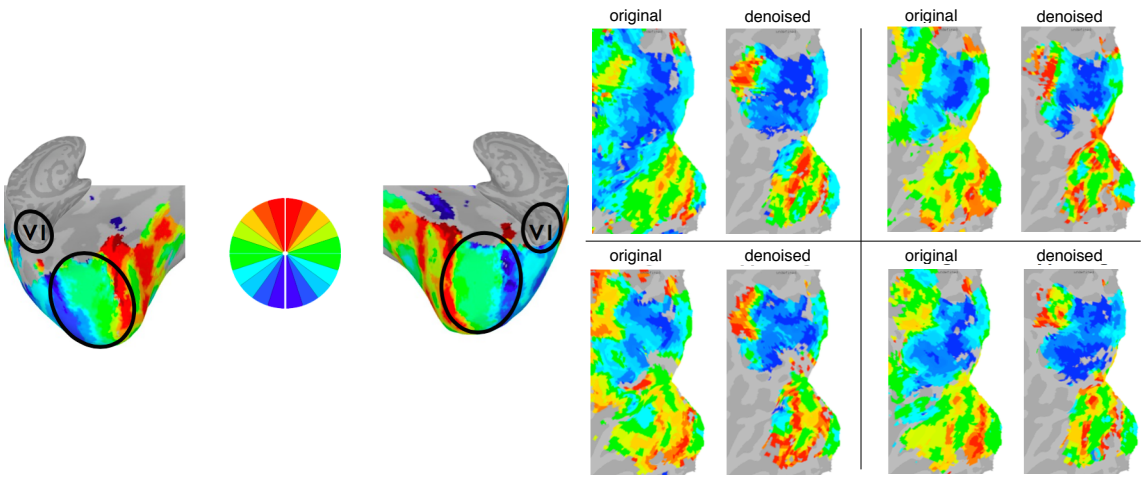


Figure 3.5: Left: Illustration of SRM with retinotopy. Right: Original and denoised retinotopy data. Figure credits: Michael J. Arcaro.

while subjects view a rotating checkerboard wedge sweeping across the visual field. The retinotopic map is colored according to the stimulus angle as in Fig. 3.5, showing that different parts of the brain tuned to specific polar angles of visual space. Since a brain map like this is highly interpretable, it can be used to verify the effectiveness of SRM through visual comparison. This is joint work with Michael J. Arcaro.

The experiment is done by training SRM on polar angle stimulus data with 18 subjects. We first plot the polar angle map without SRM in Fig. 3.5. After training, we project the data into shared feature space and then project it back to the original space. In math, this will be $X_{\text{denoised}} = WW^T X_{\text{original}}$. Both the raw polar angle map and the transformed polar angle map are plotted in Fig. 3.5 for comparison. From the figure we can observe much more vivid color contrast. A map like this with higher contrast is generally regarded as less noisy. Therefore, this suggests that the procedure has a potential to increase signal to noise ratio. At the very least, the high similarity between transformed polar angle map and the original polar angle map is a sanity check, indicating that SRM isn't breaking something terribly.

An interesting experiment is to project one subject's data into the shared feature space, and then project the feature into another subject's voxel space. The results

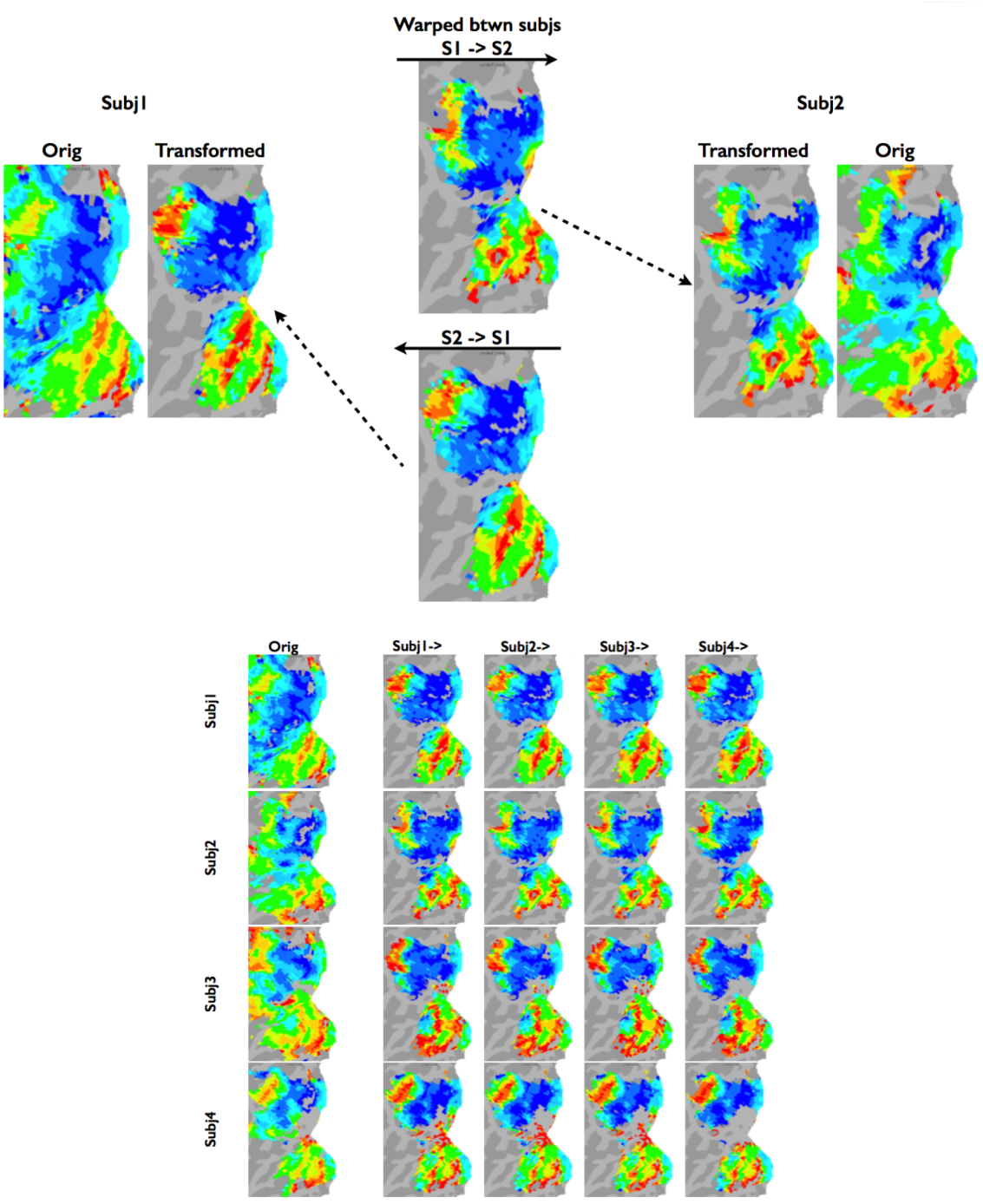


Figure 3.6: Top: Mapping data between the subject 1 and the subject 2. Bottom: Mapping data between subjects. Figure credits: Michael J. Arcaro.

of these two projections can be thought of as subject 1's brain response mapped into the subject 2's brain. This is in the same spirit as predicting the subject 2's response from the subject 1's response, shown in Fig. 3.6. The similarity between predicted

response and true response directly shows the potential of SRM for learning useful subject specific mappings that preserve informative response, in this case polar angle information. This is an example of directly demonstrating of the model's capability in learning an informative mapping without casting the verification as a prediction problem.

3.7 Searchlight Shared Response Model

The shared response model assumes a shared and temporally synchronized stimulus across subjects. Such a model can often identify shared information, but it may not be able to pinpoint with high resolution the spatial location of this information. In this experiment, we examine a searchlight based application of the shared response model to identify shared information in small contiguous regions (searchlights [77, 41]) across the whole brain. Validation using classification tasks demonstrates that we can pinpoint informative local regions. This is joint with with Hejia Zhang and has been published in [139].

A searchlight uses a fixed number of neighboring voxels to conduct analysis for each voxel location, and the same analysis is conducted on all locations. This idea can be used to extend any ROI-based factor model to overlapping searchlights. We combine factor models of the SRM form and searchlight analysis to enable localized analysis in the whole brain multi-subject fMRI analysis. In detail, a fixed sized searchlight centering at voxel i is used to scan over the whole brain. For each searchlight location, a factor model is used to functionally align across subjects, and an analysis is performed based on the results of the alignment. Statistics from the analysis (e.g., classification accuracy) is assigned to the center voxel i of the corresponding searchlight. We report the accuracy of each searchlight on a given classification task. This

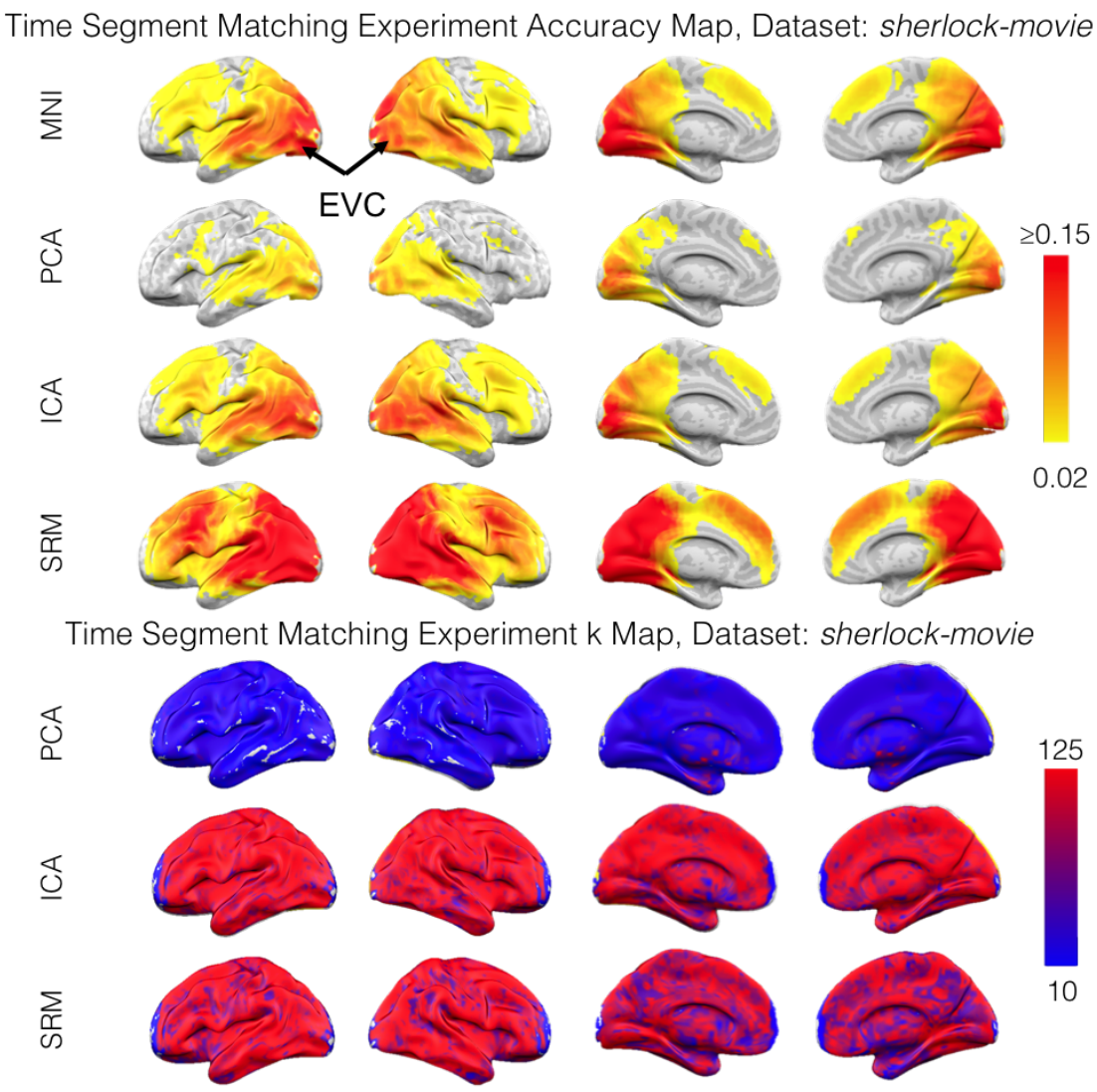


Figure 3.7: Accuracy and k brain maps for time segment matching using *sherlock-movie* dataset. EVC: early visual cortex. Figures from [139].

helps identify locations in the brain in which information is shared across subjects for a specific cognitive task.

For multi-subject neuroscience datasets and experiments, we provide an effective method for locating where the shared information is over the whole brain while keeping the quality of the found shared information. So our method can serve as a first step in multi-subject fMRI analysis to help identify regions worthy of further investigation. In some models, the number of latent factors q can be pre-specified. In this

case, we record the k value that gives the best analysis result for each searchlight. We report both the accuracy and the best k value on a brain map as a proxy for the presence and richness of a shared cognitive state across subjects.

To make things more precise, let $X_i \in \mathbb{R}^{v_x \times v_y \times v_z \times d}$ denote the data from subject i, $i = 1 : m$, where (v_x, v_y, v_z) are the number of voxels in the 3D volume along the (x, y, z) axes, and d is number of time samples (TRs) in the experiment. Typically one re-arranges X_i into a v by d matrix, where $v = v_x \times v_y \times v_z$. If one is only interested in voxels within a given region of interest (ROI), then X_i is simply the sub-matrix formed by restriction to those voxels.

Three datasets are used to test and compare the searchlight factor models: the *sherlock-movie* dataset (§3.2.2), the *sherlock-recall* dataset (§3.2.2), and the *audiobook* dataset (§3.2.3). In all experiments, we use $5 \times 5 \times 5$ searchlights on data down-sampled by 2. For each searchlight, we examine $k = [10, 25, 50, 75, 100, 125]$ and report the highest testing accuracy and the corresponding k value. The accuracies are computed based on the projected shared response of held out data using learned subject-specific maps $W_i \in \mathbb{R}^{125 \times k}$.

3.7.1 Time Segment Matching Experiment

This experiment is designed to test if the shared response we learned can be generalized to new data. That is, what is the quality of the shared information extracted. We use the *audiobook* and *sherlock-movie* datasets. The fMRI data are split into two halves along the temporal axis, one for training and the other for testing, and the roles reversed and the results averaged. In this experiment, we first use training data of all 9 subjects for learning the shared response. Then, a random 9 TR time segment from the testing subject (1 of the 9 subjects), called test segment, is projected to the shared response space. The other 8 subjects' testing data is projected to the shared response space and averaged. We then locate this time segment by

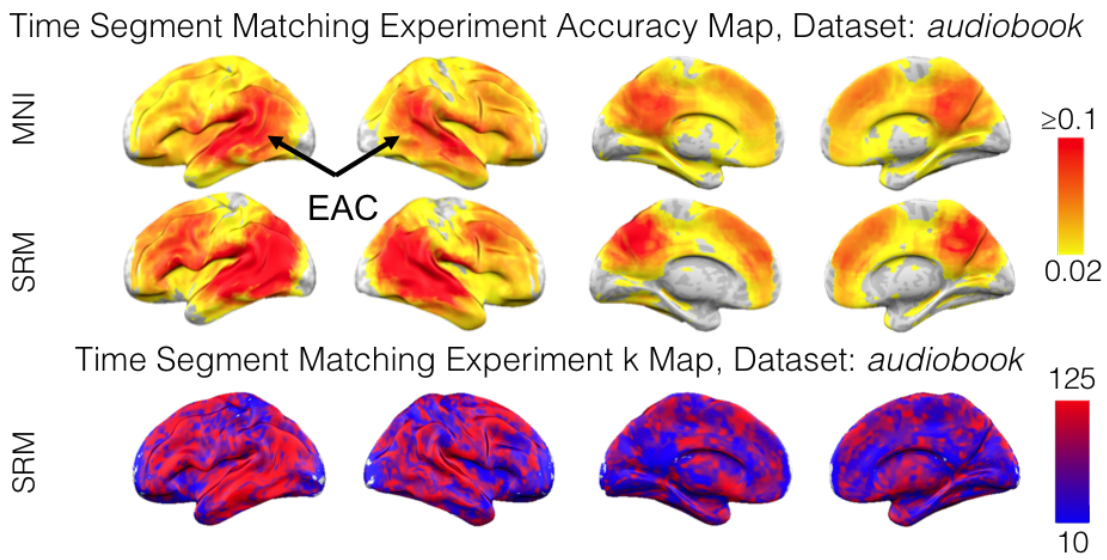


Figure 3.8: Testing accuracy and k brain maps for time segment matching experiment using audiobook dataset. EAC: early auditory cortex. Figures from [139].

maximizing Pearson correlation between the average response and response from the test segment. Segments overlapping with the test segment are excluded in matching. Assuming independent time segments, chance accuracy is 0.0044 for *audiobook* and 0.001 for *sherlock-movie*. The testing accuracy and k brain maps for different searchlight factor models are shown in in Fig. 3.7 and Fig. 3.8. Note that we threshold the accuracy to give a more clear visualization of the most informative area. We also compute a single number accuracy by aggregating the local shared response from all searchlights. This accuracy is compared with accuracy from whole brain factor models with $k = 100$ features The results are shown in Fig. 3.10.

3.7.2 Scene Recall Matching Experiment

This experiment is designed to test if brain functional patterns are similar when subjects are recalling the same scene. We use the *sherlock-movie* for training and *sherlock-recall* for testing. In *sherlock-recall*, TRs collected when the subject was recalling the same scene are averaged and projected to the shared response space

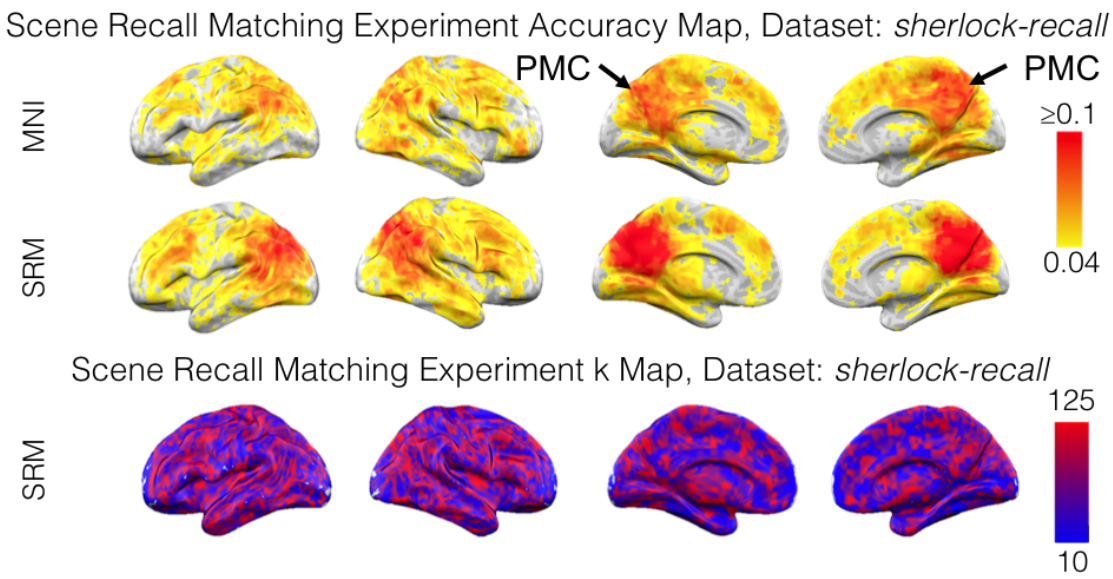


Figure 3.9: Testing accuracy and k brain maps for scene recall matching experiment using *sherlock-recall* dataset. PMC: posterior medial cortex. Figures from [139].

using W_i learned from the training data. The projected recall data along with the corresponding scene labels are used to train a SVM classifier. The projected recall data from a testing subject is used to test the classifier. Chance accuracy is 0.02. The testing accuracy and k brain maps for a subset of models are shown in Fig. 3.9. Accuracies for searchlight factor models and whole brain factor models are computed the same way as time segment matching experiment and are shown in Fig. 3.10.

We have investigated how well various factor models can locate informative regions in a searchlight based analysis of multi-subject fMRI data. This approach highlights local brain regions that are most informative of the cognitive state of interest using both accuracy and k brain maps. Early auditory cortex (EAC) and early visual cortex (EVC) are the most informative regions in time segment matching experiment for *audiobook* and *sherlock-movie* dataset, respectively. This matches the type of stimulus in these datasets. Scene recall is a more complex task. In this case a higher level cognitive region, PMC, is more informative. The results demonstrate that the approach can effectively locate meaningful informative local regions. In

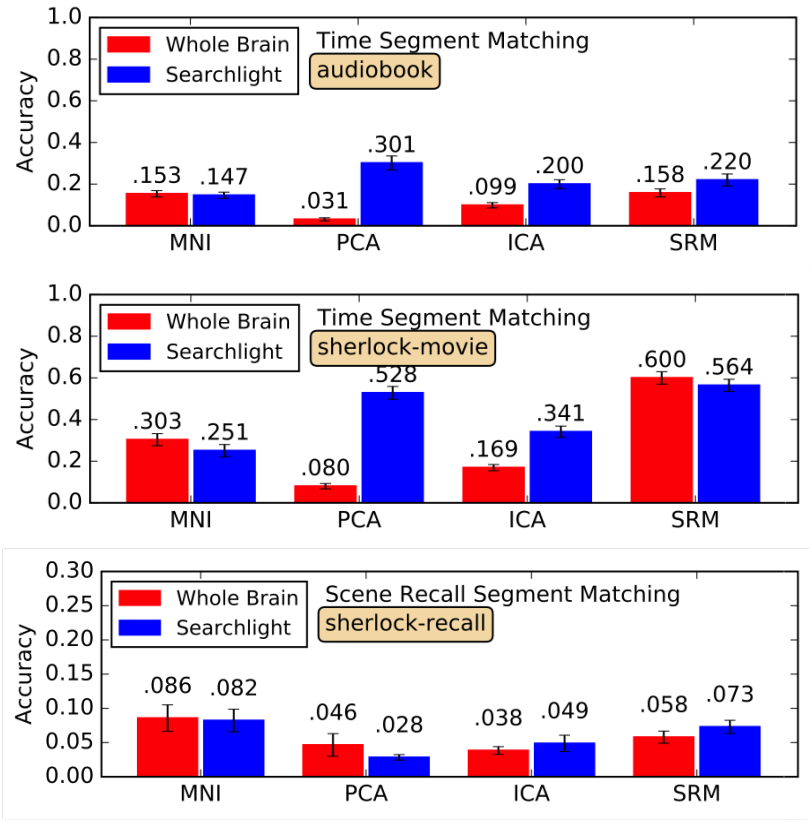


Figure 3.10: Top and Middle: Time segment matching accuracies (top: audiobook, middle: sherlock-movie). Bottom: Scene recall matching accuracy (sherlock-movie). Figures from [139].

some neuroscience experiments, it is not clear which regions will be most relevant to the stimulus and/or task. The searchlight factor model approach helps locate regions worthy of further exploration. Moreover, since the searchlight approach preserves spatial locality, we expected the overall accuracy to drop as a consequence of the searchlight constraint. In fact, as shown in Fig. 3.10, the overall accuracy does not drop in most cases, and sometimes even significantly increases.

3.8 Amount of Data Required for SRM

Future neuroimaging data collection might require a functional registration task at the beginning of scanning, just like the current practice of conducting a structural

scan for anatomical registration. For example, all subjects might be required to watch a short movie to facilitate learning SRM mappings between subjects. Since scanning time is precious, we explore the minimum amount of observations needed for learning a good shared response space. The amount of observations can be measured in two ways. One way is to measure the number of subjects needed to learn a good shared representation. Another way is to measure the number of observations needed per subject to learn a good shared representation.

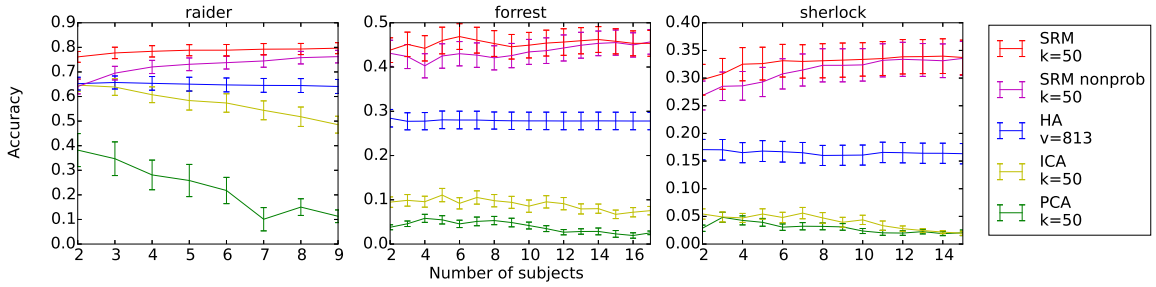


Figure 3.11: Effect of the number of subjects used in SRM training on the classification 18s time segments of a held out subject for three datasets and distinct ROIs. Error bars: ±1 stand. error.

A key question is how the number of training subjects impacts the quality of the estimated shared response and hence the matching and classification performance reported in the time segment matching experiment §3.4. In short, how many subjects do we need to learn a reliable shared template? To examine this question, we conduct the time segment matching experiment but vary the number of subjects used during the training phase. From the results (Fig. 3.13) we observe that SRM is able to learn an effective template from the data of only a few subjects and that predictive test performance increases slightly with an increase in number of training subjects. The generalization performance of the degenerate form of SRM isn't as good with only a few subjects. However, as the number of subjects increases, the degenerative form yields similar generalization performance to SRM. This indicates that the extra regularization in SRM is more helpful with less training data, which is expected. Hyperalignment performs robustly over a range of subjects. For both standard ICA

and PCA, performance decreases with the increase in the number of subjects. One possible explanation for this is that standard ICA and PCA are introducing an unnatural orthogonality constraint across the subject specific transformation matrices that hinders the methods from learning a shared response.

Another dimension is the number of TRs per subject needed to learn a good representation. We ran the image category classification experiment as in §3.4 for various numbers of TRs for fitting SRM. The results are shown as the red curve in Fig. 4.1 and in Fig. 3.12. In Fig. 4.1, we observe improved accuracy when more data is used for training SRM. At 500 TRs, SRM is able to reach significantly above chance prediction accuracy, and after 1500 TRs, the prediction accuracy plateaus. In Fig. 3.12, we also observe improved accuracy when more data is used for training SRM. However, different from the results in Fig. 4.1, we don't observe clear trend of plateaus. This might be primarily due to the maximum amount of TRs we can use for training or the different complexity in the task of testing. From these results we can see that the minimum amount of TRs required for training an SRM depends on the complexity of the training data and the complexity of the evaluation task. Empirically, 500 TRs is an adequate amount of data for training an SRM, but it may still differ from this number depends on the dataset.

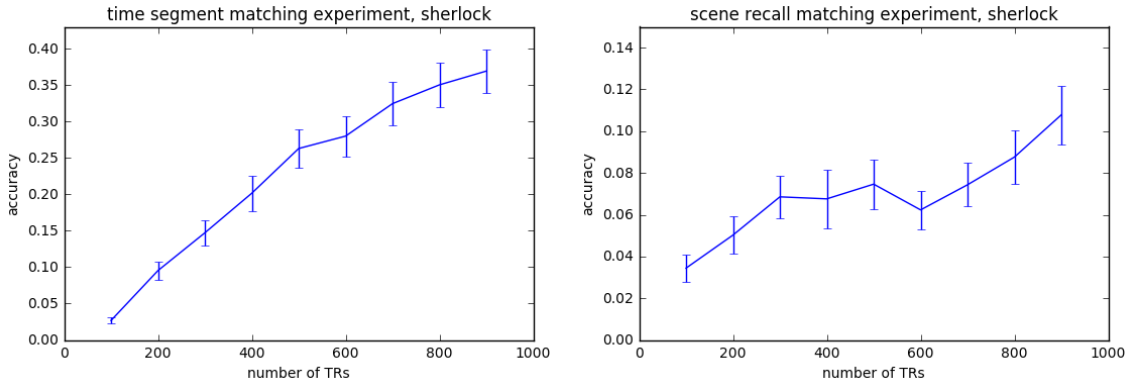


Figure 3.12: Effect of the number of TRs used in SRM training on the time segment matching experiment and scene recall matching experiment using *sherlock* dataset. Error bars: ±1 stand. error.

3.9 Other Explorations

3.9.1 Non-temporally Synchronized Stimulus

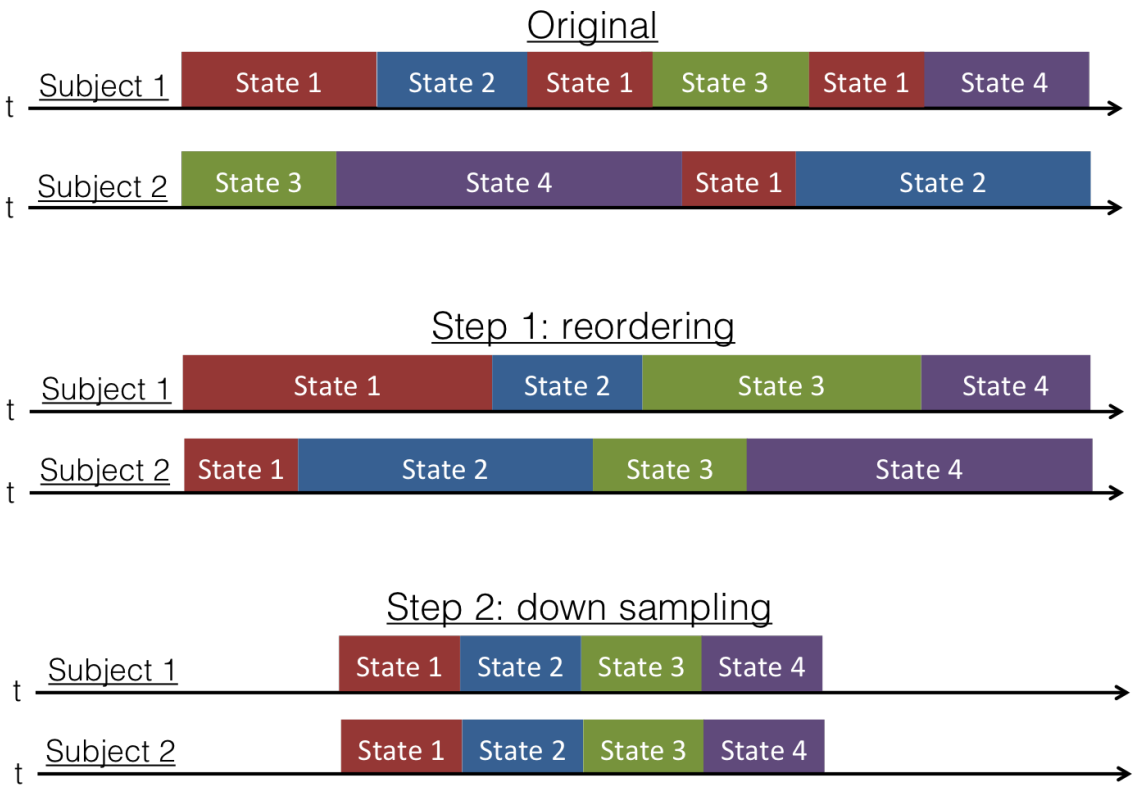


Figure 3.13: Data preprocessing steps for using SRM on non-temporally synchronized stimulus.

SRM assumes data from different views of the same timepoint are different realizations of the same underlying source. However, for many neuroimaging experiments this assumption does not hold, such as block design experiments. Can we use SRM in this regime? In this section, we explore approaches to apply SRM on non-temporally synchronized fMRI data. To do this, we assume that each observation is a noisy version of the brain state at the time. Generally, we design the experiment in a way that the brain state at a given time is well defined. For example, assume we have 2 subjects and 4 different states (shown as the top row of Fig. 3.13). The first step we can do is to reorder each subject's state transition trajectory. Since for the

same state, there might be different numbers of observations between subjects which doesn't quite fit into the SRM framework. A plausible way is to down sample the subject's state with more observations into the same number of observations as the other subject. By preprocessing the data through these approaches, now the data is compatible with SRM's assumptions. We have preliminary results demonstrating the potential of this approach. However, more experiments on more datasets are needed to draw conclusion. The robustness of this approach also requires further exploration.

3.9.2 Quantifying Shared Dimensionality

In SRM, the dimensionality k of the shared feature space is a parameter that can be adjusted. For the predictive experiment performed, different values of k lead to different predictive performance. By doing cross-validation on k, we get a value k that leads to best predictive performance in the validation set. For example, Fig. 3.3 and Fig. 3.4 shows the predictive performance altered for various values of k. Furthermore, [25] also uses SRM to quantify the dimensionality of the spatial patterns that are shared across subjects and can contribute to some classification tasks. In Fig. 3.7, Fig. 3.8, and Fig. 3.9, we plot the k brain maps for each different experiments. These figures illustrate the dimensionality of shared information across different parts of the brain using values of k selected by cross-validation as measures.

3.10 Discussion and Conclusion

The vast majority of fMRI studies require aggregation of data across individuals; in this setting, we treat data from a subject as a view. By identifying shared responses between the brains of different individuals, our model enhances fMRI analyses that use aggregated data to evaluate cognitive states. A key attribute of SRM is its built-in dimensionality reduction leading to a reduced-dimension shared feature space. We

have shown that by tuning this dimensionality, the data-driven aggregation achieved by SRM demonstrates higher sensitivity in distinguishing multivariate functional responses across cognitive states. This was shown across a variety of datasets and anatomical brain regions of interest. This also opens the door for the identification of shared and individual responses. The identification of shared responses after SRM is of great interest, as it allows us to assess the degree to which functional topography is shared across subjects. Furthermore, the SRM allows the detection of group specific responses. This was demonstrated by removing an estimated shared response to increase sensitivity in detecting group differences. We posit that this technique can be adapted to examine an array of situations where group differences are the key experimental variable. The method can facilitate studies of how neural representations are influenced by cognitive manipulations or by factors such as genetics, clinical disorders, and development.

Successful decoding of a particular cognitive state (such as a stimulus category) in a given brain area provides evidence that information relevant to that cognitive state is present in the neural activity of that brain area. Conducting such analyses in locations spanning the brain, e.g., using a searchlight approach, can facilitate the discovery of information pathways. In addition, comparison of decoding accuracies between searchlights can suggest what kind of information is present and where it is concentrated in the brain. SRM provides a more sensitive method for conducting such investigations. This may also have direct application in designing better noninvasive brain-computer interfaces [34].

Our model provides higher sensitivity in learning a shared response, but may need to be combined with other data to draw conclusions about functional relevance. One way is to test for links between behavioral performance and decoding performance. Using classification accuracy in subject viewing face or scene, a close-loop moment-to-moment feedback of attentional state to the subject has demonstrated enhancement

in attention abilities [34]. One step forward will be jointly modeling brain data with behavior data, and our probabilistic latent variable formulation opens up the potential for extensions in this direction.

Chapter 4

Extensions of SRM

4.1 Introduction

In Chapter 2, we proposed a framework for identifying a shared response based on multi-view data and evaluated the method in Chapter 3 using various neuroimaging studies. In this chapter, we explore four different extensions of the SRM framework. These extensions follow various directions, including the utilization of both labelled data and unlabeled data (§4.2), higher order statistics (§4.4, §4.3), and temporal structure (§4.5).

We start with a semi-supervised shared response model. This opens up the possibility of using both labelled data and unlabeled data to estimate a shared response. We then introduce a shared response ICA. This alters the SRM objective function from maximizing variance to maximizing independence. These two extensions are done with collaborators. We then introduce a kernelized version of the shared response model. This allows utilization of higher order statistics into the model. Lastly, we introduce a gaussian process shared response model. This models the latent variables as a gaussian process.

Prior Publications and Acknowledgment Parts of §4.3 have been published in [139], and parts of §4.2 have been published in [120]. These are collaborative efforts. In §4.3, my primary contribution is the design of algorithms, initial implementation of the algorithms, and the design of the experiments. In §4.2, my primary contribution is the design of algorithm and the design of the experiments. For parts that aren't my primary contribution, I'll briefly describe them in this chapter for continuity and self-containment purposes, but will refer to the corresponding papers for further details.

I thank Hejia Zhang for her permission to use figures from [139] in §4.3, Jacob Simon for his help in coding SR-ICA in §4.3, and Javier S. Turek for his permission to use figures from [120] in §4.2.

4.2 Semi-supervised Shared Response Model

In a typical experiment for evaluating the performance of the fitted model, we train the model on unlabeled data and use the model to test prediction on labeled data. The reason for this two phase approach is because we want to test if the trained model generalizes to new data and new subject. However, depending on the application that we are interested in, we might want to achieve the best predictive performance as possible. This might be the case in an application to real-time fMRI or a brain-computer interface. Under these scenarios, a better approach is to combine both labeled and unlabeled datasets during the training phase. This motivates us to design a semi-supervised version of shared response model. This is joint work with Javier S. Turek.

4.2.1 Mathematical Formulation

We propose to jointly learn subject specific basis for projecting to shared feature space and a classifier in the shared feature space. We use a multinomial logistic regression

as the classifier for deriving the model. However the framework allows for different types of classifiers.

We design the model by combing the loss function of unsupervised SRM, $\mathcal{L}_{Align}$, and the loss $\mathcal{L}_{Sup}$ of a supervised classifier. We require the model to simultaneously learn a low dimensional data representation as well as a classifier that classifies well on the low dimensional representation of labelled data. The proposed semi-supervised scheme involves solving

$$\min_{\psi,\theta} (1-\alpha)\,\mathcal{L}_{Align}(\psi) + \alpha\mathcal{L}_{Sup}(\theta;\psi) + R(\theta), \tag{4.1}$$

where ψ and θ are the semi-supervised model parameters, $R(\theta)$ is a regularization term for the supervised task, and $\alpha \in [0,1]$ is a scalar value that controls the bias between the functional alignment term and the supervised term. In solving (4.1), we learn the transformation parameters and supervised classifier jointly. This is different from the traditional approach of first learning the transformation parameters in an unsupervised way and then learning a supervised classifier. The semi-supervised approach allows the information from supervised learning to aid the optimization of the unsupervised tasks, while the traditional approach doesn't allow feedback from the supervised task to effect the unsupervised task. In (4.1), we also introduce a parameter α to control the balance between the unsupervised task and the supervised task. When $\alpha = 0$, the model only focuses on the unsupervised information, and for $\alpha = 1$, the model only focus on the supervised information.

4.2.2 Experiment

We evaluate the proposed semi-supervised framework using classification experiments. The specific instantiated model from the semi-supervised framework is a combination of SRM and multinomial logistic regression (MLR), which we call semi-supervised

SRM (SS-SRM). Please refer to [120] for detailed derivation of the inference algorithm for SS-SRM.

Two datasets are used for the experiment, the *raider* dataset (§3.2.1) and the *sherlock* dataset (§3.2.2). The movie data from *raider* and *sherlock* are used to train the unsupervised part of the framework, respectively. The image data from *raider* with category labels and the recall data from *sherlock* with scene labels are used to train the unsupervised part of the framework, respectively. The SS-SRM is implemented in Python using the pyManOpt package [119] for updating the mappings using the Conjugate Gradient method [110] on the Stiefel manifold [37]. The number of iterations is fixed to be 15 for both SRM and SS-SRM.

In the first experiment, we conduct the image classification experiment with the *raider* dataset (§3.2.1) as in §3.4 with different numbers of observations for the unsupervised training part. By varying the number of observations, we can evaluate how much unlabeled data is needed to reach an acceptable predictive performance. We run the experiment with three different methods: a plain MLR classifier without unsupervised training, SRM followed by an MLR classifier, and SS-SRM. All methods use an l_2-regularization. The regularizer parameter for the MLR classifier is $\gamma = 0.001$. SRM was trained with $k = 50$ features for the shared response and the regularizer value for MLR is $\gamma = 0.001$. The SS-SRM method also used $k = 50$ features and parameters $\alpha = 0.2$ and $\gamma = 1.0$. The average accuracy performance of the methods and their standard errors are presented in Figure 4.1. The resulting predicting performance including the entire movie stimuli is also shown in Table 4.1.

We observe significant improvement with the SRM over a plain MLR classifier, while SS-SRM reaches better accuracy than SRM for any number of unsupervised training observations. SS-SRM requires about half the number of observations from the unsupervised task to achieve the same results as SRM using the whole movie. This demonstrates the potential for a semi-supervised method to reach the same level

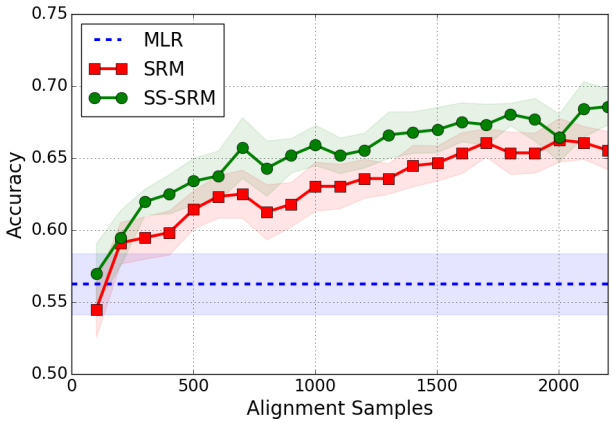


Figure 4.1: Average accuracy as a function of the number of training samples. Figure from [120].

Dataset	Experiment	MLR	SRM	SS-SRM
raider	Image category	56.25%	65.53%	68.57%
sherlock	Scene recall	4.28%	5.31%	6.12%

Table 4.1: Comparison of average accuracy for brain decoding experiments.

of accuracy while using far less fRMI observations. It also shows improvement in prediction performance by using more unlabeled data. This is a critical property for semi-supervised method because it allow us to use large unlabeled dataset rather than a large labeled dataset. This is cheaper and and easier to collect.

The second experiment we conduct is the scene recall experiment with the *sherlock* dataset (§3.2.2) as in §3.7. The three methods: plain MLR, SRM followed by MLR, and SS-SRM are tested. The MLR classifier used $\gamma = 0.01$, SRM with MLR ran with $\gamma = 10$, whereas SS-SRM uses $\gamma = 0.1$. Both, SRM and SS-SRM used $k = 25$ features for the shared response dimension. Table 4.1 shows the accuracy for these methods. The low prediction accuracy is primarily due to the difficulty of this task, however, all methods are above the chance level of 2.12%[1]. SS-SRM achieves better predictive accuracy than all the other methods.

[1]Three scenes were recalled only by less than 3 subjects and they were removed from the data, leaving 47 different scenes for the experiment.

4.2.3 Discussion and Conclusion

In this section, we present a semi-supervised framework allowing joint learning of a multi-view representation model with unlabeled data and a classifier with labeled data. The integration of the unsupervised task and the supervised task allows feedback of the error to effect both sub-tasks. By combining the shared response model (SRM) and multinomial logistic regression (MLR), we obtained an instantiation of the semi-supervised framework, which we called SS-SRM. SS-SRM achieves higher predictive accuracy than using SRM followed by an independent MLR. We also find that the same accuracy of SRM can be achieved with SS-SRM while using less input data. This is critical in neuroscience research for two reasons. First, there is a natural limitation of how many observations can we gather within one session. Second, subjects get tired over time, so the best data may be at the beginning of the experiment. By reducing the number of observations needed for the multi-view learning part, it allows more time to conduct additional neuroscientific experiments.

4.3 Shared Response Independent Component Analysis

In §2.2.1, we demonstrate that in essence SRM is maximizing within view variance and sum of pair-wise covariance. In this section, we explore different objective functions under the same framework. Specifically, we develop two new variants of the ICA and group ICA factor models. These variants also show good performance in the functional alignment task.

Algorithm 1: Shared Response ICA (SR-ICA)

input : Data matrices X_i, number of factors k, convergence threshold τ, max iteration N, number of subjects m

output: Subject-specific maps W_i and shared response S

$W_i^0 \leftarrow$ initialization with random orthonormal columns ;
for n *in* 1 *to* N **do**
  $S \leftarrow \frac{1}{m}\sum_{i=1}^{m} W_i^{n-1^+} X_i$ ▷ $(\cdot)^+$ is pseudo-inverse;
  for i *in* 1 *to* m **do**
    $W_i^n \leftarrow (E\{X_i g(S)\} - E\{X_i g'(S)\} W_i^{n-1^+})^+$;
    $W_i^n \leftarrow W_i^n (W_i^{nT} W_i^n)^{-1/2}$;
  end
  converged $\leftarrow True$;
  for i *in* 1 *to* m **do**
    if $max\ |W_i^{nT} W_i^{n-1} - I| \geq \tau$ **then**
      converged $\leftarrow False$;
    end
  end
  return W_i, S;
end

Algorithm 2: Shared Response Group ICA (SR-GICA)

input : Data matrices X_i, number of factors k_1, k_2, number of subjects m

output: Subject-specific maps W_i and shared response S

for i *in* 1 *to* m **do**
  $X_i = F_i P_i$ ▷ First PCA with k_1 components;
end
$P \leftarrow [P_1^T, \ldots, P_m^T]^T$;
$P = GY$ ▷ Second PCA with k_2 components;
$Y = AS$ ▷ ICA with k_2 components;
Partition $[G_1^T, \ldots, G_m^T]^T \leftarrow G$;
Then, $G_i AS = P_i \rightarrow F_i G_i AS = F_i P_i = X_i$;
$W_i \leftarrow F_i G_i A$;
return W_i, S;

4.3.1 Mathematical Formulation

A desirable factor model should have the following properties: 1) have an adjustable parameter selecting the number of factors k; 2) exhibit good performance in large area multi-subject functional alignment.

We start with reviewing ICA under the shared response framework:

Independent Component Analysis (ICA) ICA learns statistically independent signals as measured by kurtosis or negentropy [67, 80]. We use the FastICA algorithm. This is an efficient probabilistic method [67] optimizing negentropy of the shared response S. This is formulated as

$$\max_W [E(G(S)) - E(G(\mathcal{N}))]^2, \tag{4.2}$$

where $G(\cdot)$ is a nonquadratic function, e.g. $\log\cosh$, and $\mathcal{N}$ is a standard normal random variable. This yields $X = WS + E$. W is then partitioned into m submatrices W_i, $i = 1 : m$.

Shared Response Independent Component Analysis (SR-ICA) We propose a new algorithm call shared response ICA (SR-ICA) by modifying the FastICA algorithm. It is motivated by the framework in §2. In SR-ICA, the block structure of the subject data is preserved by spatial concatenation in both X and W. The key difference is that instead of learning a joint matrix W, we iteratively learn W_i to ensure block-wise structure in W. This is summarized in Algorithm 1[2]. Here we follow the convention of working with unmixing matrices U_i instead of W_i. The function $g(\cdot)$ is the derivative of $G(\cdot)$ in (4.2) [68].

[2]We acknowledge the help of Jacob Simon in coding SR-ICA.

Group ICA (GICA) Group ICA, an algorithm for making group inferences, uses two applications of PCA and an application of ICA [20]. The original algorithm first performs subject specific PCA along the temporal dimension. Then the projected data matrices for all subjects are concatenated to form a joint data matrix. A second PCA is then performed on the joint data matrix. Lastly, an ICA is performed on the projected data matrix after the second PCA. We apply GICA along the spatial dimension to learn a low dimensional shared response space. See Algorithm 2.

4.3.2 Experiments and Results

Three datasets are used in this section, including *audiobook* (§3.2.3), *sherlock-movie* and *sherlock-recall* (§3.2.2). In all experiments, we use $5 \times 5 \times 5$ searchlights on data down-sampled by 2. For each searchlight, we try $k = [10, 25, 50, 75, 100, 125]$ and report the highest accuracy and the corresponding k value. To ensure a fair comparison, for SR-GICA, we set k_1 to be the number of voxels in the searchlight and $k_2 = k$. We tried other k_1 values, but this resulted in lower accuracy. Note that there are relatively few voxels per searchlight to begin with. The accuracies are computed based on the projected shared response of held out data using learned subject-specific maps $W_i \in \mathbb{R}^{125 \times k}$.

Time Segment Matching This experiment is designed to test if the shared response we learned generalizes to new data. That is, what is the quality of the shared information extracted. We use the *audiobook* and *sherlock-movie* datasets. The fMRI data are split into two halves along the temporal axis, one for training and the other for testing, and the roles reversed and the results averaged. In this experiment, we first use training data of all 9 subjects for learning the shared response. Then, a random 9 TR time segment from the testing subject (1 of the 9 subjects), called test segment, is projected to the shared response space. The other 8 subjects' testing

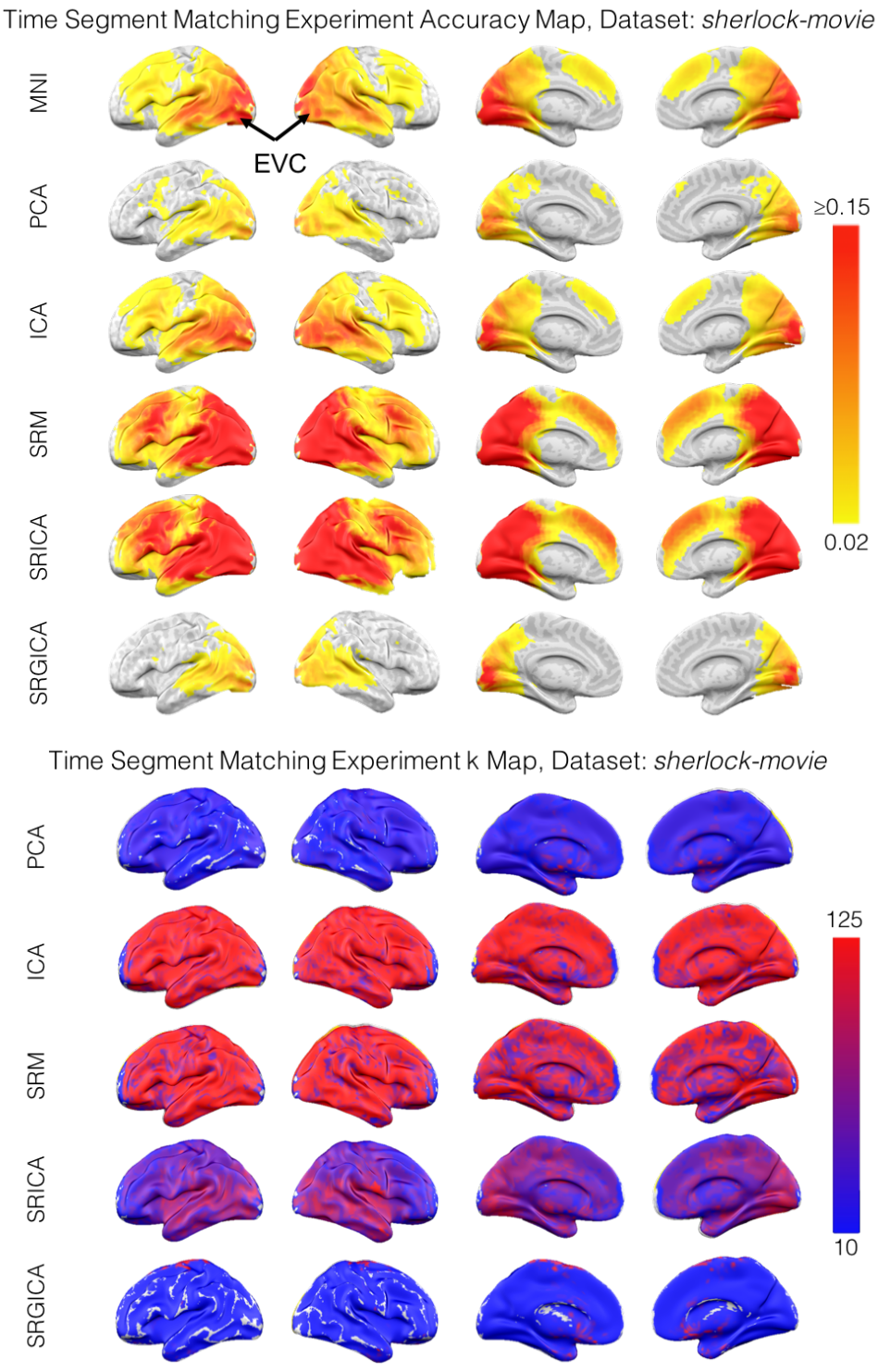


Figure 4.2: Top: Accuracy and k brain maps for time segment matching using *sherlock-movie* dataset. EVC: early visual cortex. Figures from [139].

data is projected to the shared response space and averaged. We then locate the testing time segment by maximizing Pearson correlation between the average response and response from the test segment. Segments overlapping with the test segment are excluded in matching. Assuming independent time segments, chance accuracy is 0.0044 for *audiobook* and 0.001 for *sherlock-movie*. The accuracy and k brain maps for different searchlight factor models are shown in Fig. 4.2 and Fig. 4.3. Note that we threshold the accuracy to give a more clear visualization of the most informative area. We also compute a single number accuracy by aggregating the local shared response from all searchlights. This accuracy is compared with accuracy from whole brain factor models with $k = 100$ features ($k_1 = 500, k_2 = 100$ for SR-GICA). The results are shown in Fig. 4.5.

Scene Recall matching This experiment is designed to test if brain functional patterns are similar when subjects are recalling the same scene. We use the *sherlock-movie* for training and *sherlock-recall* for testing. In *sherlock-recall*, TRs collected when the subject was recalling the same scene are averaged and projected to the shared response space using W_i learned from the training data. The projected recall data along with the corresponding scene labels are used to train a SVM classifier. The projected recall data from a testing subject is used to test the classifier. Chance accuracy is 0.02. The accuracy and k brain maps for a subset of models are shown in Fig. 4.4. Accuracies for searchlight factor models and whole brain factor models are computed the same way as time segment matching experiment and are shown in Fig. 4.5.

4.3.3 Discussion and Conclusion

We have investigated how well SRICA and SRGICA can locate informative regions in a searchlight based analysis of multi-subject fMRI data. This approach highlights

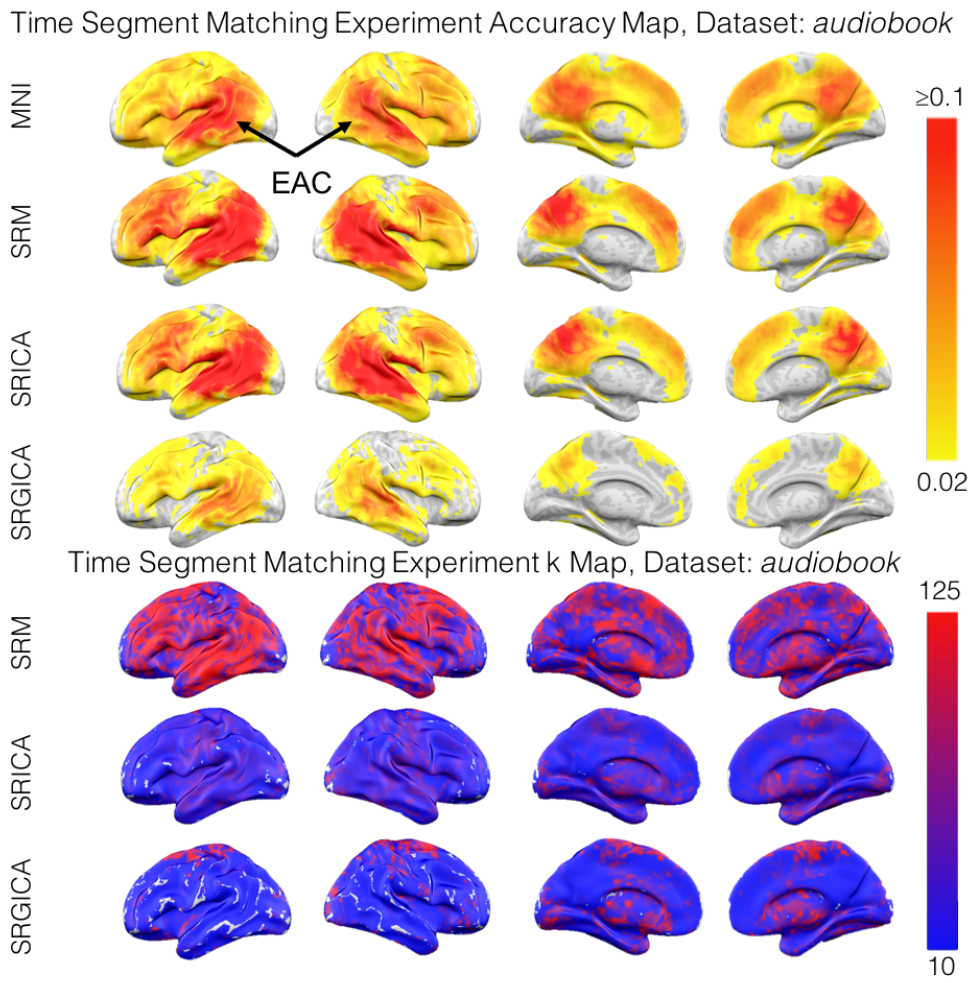


Figure 4.3: Accuracy and k brain maps for time segment matching experiment using *audiobook* dataset. EAC: early auditory cortex. Figures from [139].

local brain regions that are most informative of the cognitive state of interest using both accuracy and k brain maps. Early auditory cortex (EAC) and early visual cortex (EVC) are the most informative regions in time segment matching experiment for the *audiobook* and *sherlock-movie* dataset, respectively. This matches the type of stimulus in these datasets. Scene recall is a more complex task. In this case a higher level cognitive region, PMC, is more informative. The results demonstrate that the approach can effectively locate meaningful informative local regions. In some neuroscience experiments, it is not clear which regions will be most relevant to the stimulus and/or task. Using SR-ICA and SR-GICA in a searchlight setting

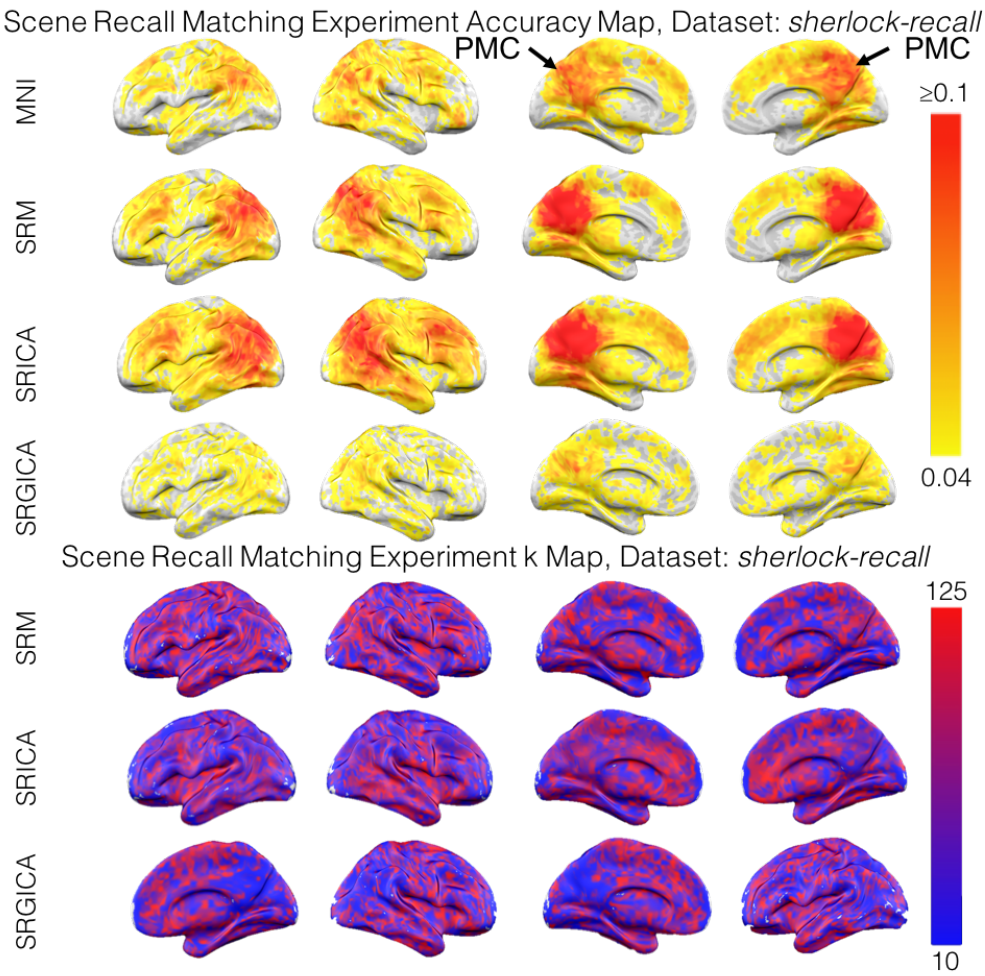


Figure 4.4: Accuracy and k brain maps for scene recall matching experiment using *sherlock-recall* dataset. PMC: posterior medial cortex. Figures from [139].

helps locate regions worthy of further exploration. Moreover, since the searchlight approach preserves spatial locality, we expected the overall accuracy to drop as a consequence of the searchlight constraint. In fact, as shown in Fig. 4.5, the overall accuracy does not drop in most cases, and sometimes significantly increases. The k brain maps also has the potential to help reveal the effectiveness of the various factor models. For example, consider the k brain maps of SR-ICA and SRM. While the accuracy maps of these methods are very close to each other, SR-ICA uses a smaller k to achieve this accuracy. This suggests that each factor in SR-ICA is more likely

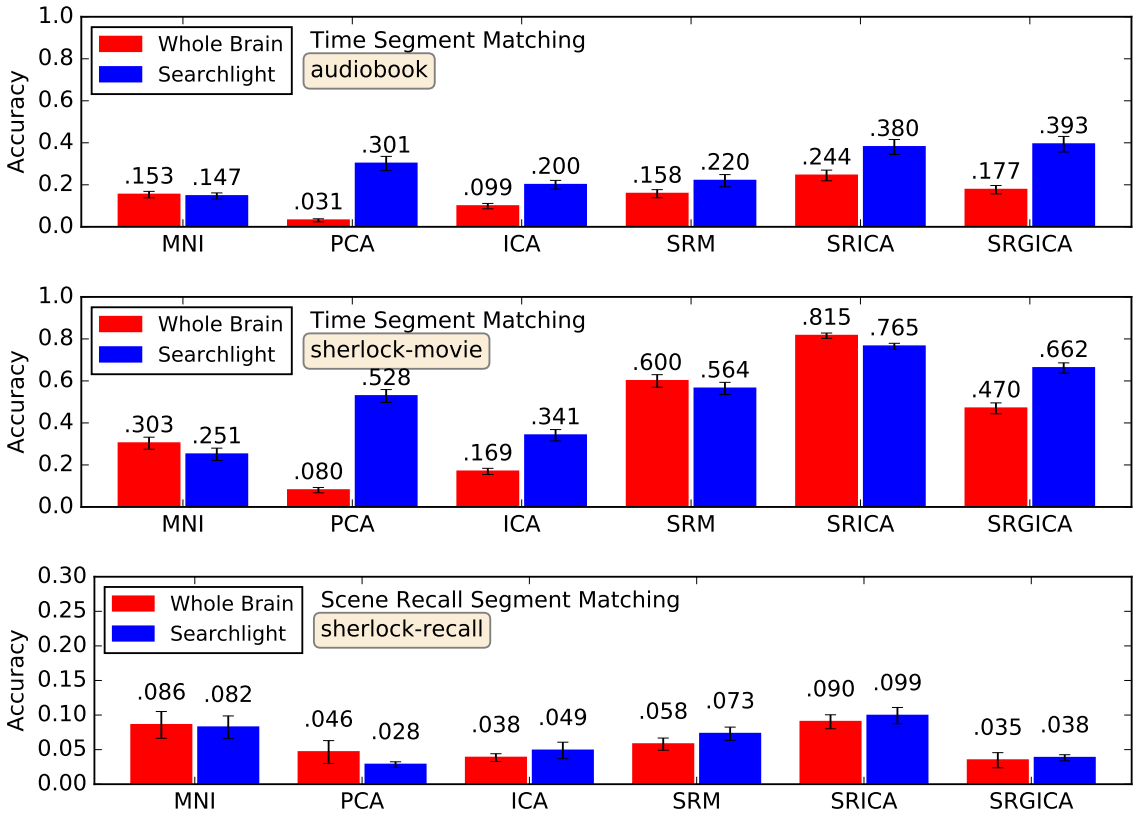


Figure 4.5: Top and Middle: Time segment matching accuracies (top: audiobook, middle: sherlock-movie). Bottom: Scene recall matching accuracy (sherlock-movie). Figures from [139].

to have an improved representation capability. Overall the SR-ICA is showing strong performance across the three experiments.

4.4 Kernelized Shared Response Model

In the SRM framework, we learn subject specific basis which can be viewed as a set of first order spatial brain maps as functional topographies. However, it is known that different parts of the brain work with each other. For example, research on functional connectivity [45, 112, 51] investigates second order statistic of the brain data. With the goal of investigating the possibility of learning higher order brain patterns, we extend the SRM into a kernelized formulation. By using the kernel trick

in our computation, we are able to explore a wider variety of factor models without incurring much computation.

4.4.1 Mathematical Formulation

Let's start with the most basic formulation of SRM as in (2.1):

$$\begin{aligned} &\min \|X_i - W_i S\|_F^2 \\ &\text{s.t.} W_i^T W_i = I, \end{aligned}$$

A natural way to extend this model is by introducing feature mapping Φ into the formulation: $\Phi_i = \Phi(X_i) \in \mathbb{R}^{N \times d}$, where N is the dimension of the feature space. $\tilde{W} \in \mathbb{R}^{N \times k}$, and $\tilde{S} \in \mathbb{R}^{k \times d}$. Rewriting the model by mapping the data X_i through feature mapping Φ, we get

$$\begin{aligned} &\min \|\Phi_i - \tilde{W}_i \tilde{S}\|_F^2 \\ &\text{s.t.} \tilde{W}_i^T \tilde{W}_i = I_k \end{aligned} \tag{4.3}$$

From representer theorem [73, 109], we know that the optimal $\tilde{W}$ lies in the span of Φ. Therefore, by letting $\mathbf{K}_i = \Phi_i^T \Phi_i \in \mathbb{R}^{d \times d}$ and parametrizing $\tilde{W}_i = \Phi_i \tilde{A}_i$, where $\tilde{A}_i$ is a generic matrix that needs to be learned, (4.3) can be rewritten as

$$\begin{aligned} &\min \|\Phi_i - \Phi_i \tilde{A}_i \tilde{S}\|_F^2 \\ &\text{s.t.} \tilde{A}_i^T \mathbf{K}_i \tilde{A}_i = I_k. \end{aligned} \tag{4.4}$$

Expanding the objective function of (4.4), we obtain

$$\begin{aligned}
&\|\Phi_i - \Phi_i \tilde{A}_i \tilde{S}\|_F^2 \\
=&\mathrm{tr}\big((\Phi_i - \Phi_i \tilde{A}_i \tilde{S})^T (\Phi_i - \Phi_i \tilde{A}_i \tilde{S})\big) \\
=&\mathrm{tr}\big(\Phi_i^T \Phi_i - 2\Phi_i^T \Phi_i \tilde{A}_i \tilde{S} + \tilde{S}^T \tilde{S}\big) \\
=&\mathrm{tr}\big(\mathbf{K}_i - 2\mathbf{K}_i \tilde{A}_i \tilde{S} + \tilde{S}^T \tilde{S}\big). \qquad (4.5)
\end{aligned}$$

By taking the derivative of the objective function with respect to $\tilde{S}$ and setting this equal to zero, we obtain the update equation

$$\tilde{S} = \frac{1}{m} \sum_i \tilde{A}_i \mathbf{K}_i.$$

To estimate $\tilde{A}_i$, we solve the following optimization problem,

$$\begin{aligned}
&\max \ \mathrm{tr}(\mathbf{K}_i \tilde{A}_i \tilde{S}) \\
&\text{s.t.} \tilde{A}_i^T \mathbf{K}_i \tilde{A}_i = I_k.
\end{aligned}$$

This is reduced from problem (4.4) by removing the first and third term in equation (4.5) because they do not depend on $\tilde{A}_i$. Then by using the Lagrangian method, we have

$$\begin{aligned}
&\max \ \ l = \mathrm{tr}(\mathbf{K}_i \tilde{A}_i \tilde{S}) + \mathrm{tr}(\Lambda_i (I - \tilde{A}_i^T \mathbf{K}_i \tilde{A}_i)) \\
&\frac{\partial l}{\partial \tilde{A}_i} = 0 \Rightarrow \mathbf{K}_i \tilde{S}^T = 2\mathbf{K}_i \tilde{A}_i \Lambda_i \\
&\frac{\partial l}{\partial \Lambda_i} = 0 \Rightarrow \tilde{A}_i^T \mathbf{K}_i \tilde{A}_i = I. \qquad (4.6)
\end{aligned}$$

Assuming $\mathbf{K}_i$ to be positive definite, this yields $A_i = \frac{1}{2}\tilde{S}^T\Lambda^{-1}$. Substituting this into (4.6), we obtain

$$\tilde{A}_i = \tilde{S}^T Q_i \Sigma_i^{-\frac{1}{2}} Q_i^T$$
$$Q_i \Sigma_i Q_i^T = \text{SVD}(\tilde{S}\mathbf{K}_i\tilde{S}^T).$$

The original SRM iterates between updating W_i and S. Estimating W_i requires taking SVD of $X_i S^T \in \mathbb{R}^{v\times k}$, with complexity $O(vk^2)$. In kernel SRM, it iterates between updating $\tilde{A}_i$ and $\tilde{S}$. Updating $\tilde{A}_i$ requires taking SVD of $\tilde{S}\mathbf{K}_i\tilde{S}^T \in \mathbb{R}^{k\times k}$, with complexity $O(k^3)$. Although the kernel matrix $\mathbf{K}_i$ is needed, it only needs to be calculated once at the beginning.

KSRM with linear kernel recovers SRM KSRM with a linear kernel recovers the original SRM. To show this we start with the update equations of SRM and Kernel SRM. For SRM we have

$$S = \frac{1}{m}\sum_i W_i^T X_i$$
$$W_i = U_i V_i^T$$
$$\text{where } U_i \Omega_i V_i^T = \text{SVD}(X_i S^T). \tag{4.7}$$

For Kernel SRM we have

$$S = \frac{1}{m}\sum_i \tilde{A}_i^T \mathbf{K}_i$$
$$\tilde{A}_i = \tilde{S}^T Q_i \Sigma_i^{-\frac{1}{2}} Q_i^T$$
$$\text{where } Q_i \Sigma_i Q_i^T = \text{SVD}(\tilde{S}\mathbf{K}_i\tilde{S}^T). \tag{4.8}$$

We show that the two sets of update equations are identical when a linear kernel, $\mathbf{K}_i = X_i^T X_i$, is used. By replacing $\mathbf{K}_i = X_i^T X_i$ into (4.8), we get

$$\begin{aligned}\tilde{S} &= \frac{1}{m}\sum_i \tilde{A}_i^T \mathbf{K}_i \\ &= \frac{1}{m}\sum_i \tilde{A}_i^T X_i^T X_i \\ &= \frac{1}{m}\sum_i W_i X_i,\end{aligned}$$

and

$$\begin{aligned}\mathrm{SVD}(\tilde{S}\mathbf{K}_i\tilde{S}^T) =& \mathrm{SVD}(\tilde{S}X_i^T X_i \tilde{S}^T) \\ =& \mathrm{SVD}((V_i\Omega_i U_i^T)(U_i\Omega_i V_i^T)) \\ =& \mathrm{SVD}(V_i\Omega_i^2 V_i^T). \qquad (4.9)\end{aligned}$$

Comparing (4.8) and (4.9), we get $Q_i = V_i$ and $\Sigma_i = \Omega_i^2$. Then by replacing these into the update equations for $\tilde{W}_i$, we get

$$\begin{aligned}\tilde{W}_i =& \Phi_i \tilde{A}_i = X_i \tilde{A}_i \\ =& X_i \tilde{S}^T Q_i \Sigma_i^{-\frac{1}{2}} Q_i^T \\ =& U_i \Omega_i V_i^T V_i \Omega_i^{-1} V_i^T \\ =& U_i V_i^T.\end{aligned}$$

This is equivalent to the solution of (2.1).

4.4.2 Discussion and Conclusion

It is commonly assumed that there is additional information in the higher order statistics of fMRI data. In this section, we developed a kernelized shared response model (KSRM) for incorporating these higher order statistics. KSRM uses the kernel trick to allow the model to incorporate higher order statistics without extra computation. The kernel trick also reduces the computation bottleneck to $\mathcal{O}(k^3)$ given the Gram matrix. This is a one-time computational cost. We have also shown that with a linear kernel, KSRM reduces to the original SRM formulation. Therefore, by simply using the kernel trick, we can gain in computation performance.

However, after trying KSRM with various different kernels (square exponential kernel, periodic kernel, local periodic kernel, linear kernel, spectral mixture kernel and combinations of these basic kernels through addition or multiplication), we haven't observed consistent statistically significant improvement over SRM in the experiments in §3.4. In most of the cases, it maintains similar predictive performance, and in some cases, the performance drops significantly. There are several factors that might lead to these results. First, it is not clear that there is stable higher order information in fMRI data. It's possible that higher order information isn't stable. This makes it much harder to estimate. Second, even with stationary higher order statistics, it is also unclear whether we have enough data to obtain a robust estimate. Third, there might be large intrinsic noise within subject such that all regions are correlated. This makes it hard to detect the relatively weak functional connectivity signal. All of these directions need to be further explored.

4.5 Gaussian Process Shared Response Model

One of the key assumptions we made in the SRM framework is temporal independence. However, it is known that various types of temporal structure exist in the

data. First, the haemodynamic response [64] of neuron stimulation last around 4 to 6 seconds [64]. Since the temporal resolution for MRI machine is generally faster than the haemodynamic response, the haemodynamic response lasts over multiple TRs. This leads to temporal structure in observed fMRI data. Second, there are intrinsic temporal structures in the stimulus. Naturalistic stimuli and block designed experiments have strong temporal structure. Third, we also expect temporal structure in the underlying brain mechanism. This leads us to propose a gaussian process shared response model (GP-SRM). This proposal explicitly models the temporal dependence using a Gaussian process.

4.5.1 Mathematical Formulation

In GP-SRM we model temporal structure by imposing a Gaussian process prior over each element of $\mathbf{s}_t$.To do so we use a zero mean and a time dependent kernel. The ith element of $\mathbf{s}_t$ is a T-dimension vector $\mathbf{s}_{ri}$ with prior $\mathcal{GP}(0, \mathbf{K}_{\mathbf{s}_{ri}}(t, t'))$. The observation for dataset m then has the form $\mathbf{x}'_{mt}|\mathbf{s}_t \sim \mathcal{N}(W_m\mathbf{s}_t + \mu_m, \Sigma_{\mathbf{x}'_m})$, where $\mathbf{x}'_{mt} \in \mathbb{R}^d$, and $W_m \in \mathbb{R}^{d\times k}$, $m = 1{:}M$. In this model, $\mathbf{s}_{ri}$ are time series factors, and the observed time series for each voxel is assumed to be a linear combination of these factors. Through the transformation matrices W_m, we hope to learn temporally structured latent time series factors. The model can be written as

$$\begin{aligned} \mathbf{s}_{ri} &\sim \mathcal{GP}(0, \mathbf{K}_{\mathbf{s}_i}(t, t')), \\ \mathbf{x}'_{mt}|\mathbf{s}_t &\sim \mathcal{N}(W_m\mathbf{s}_t + \mu_m, \rho_m^2 I), \\ \text{s.t. } W_m^T W_m &= I, \\ [\mathbf{s}_{r1} \dots \mathbf{s}_{ri} \dots \mathbf{s}_{rN}]^T &= [\mathbf{s}_1 \dots \mathbf{s}_t \dots \mathbf{s}_T], \end{aligned} \tag{4.10}$$

where $\mathbf{K}_{\mathbf{s}_i}(\cdot)$ is the kernel function.

for *i = 1 to K* **do**
  $\mathbf{s}_{ri} \sim \mathcal{GP}(0, \mathbf{K}_{\mathbf{s}_i}(t, t'))$
end
for *t = 1 to T* **do**
  for *m = 1 to M* **do**
    Generate subject specific observations:
      $\mathbf{x}'_{mt} \sim \mathcal{N}(W_m \mathbf{s}_t + \mu_m, \rho_m^2 I)$
  end
end

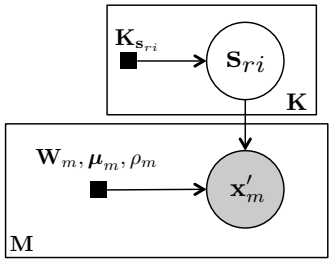


Figure 4.6: **Generative process (left) and graphical model (right) for GP-SRM.** Brain activation pattern $X_m \in \mathbb{R}^{N \times T}$ (N voxels, T TRs) is observed from subject m, $m = 1 : M$. Each column of observations X_m is a linear combination of subject specific orthogonal brain responses (columns of W_m) using the weights specified by $\mathbf{s}_t$, ith element of $\mathbf{s}_t$ is the tth element of $\mathbf{s}_i$. Shaded nodes: observations, and black squares: hyperparameters.

4.5.2 Variational Inference

For SRM, we derive a constrained EM algorithm for parameter estimation leading to (local) maximum likelihood solutions. However, due to the nonparametric nature of a Gaussian process, the same technique can't be applied. Here we derive a mean field variational inference for GP-SRM. Variational inference finds an approximate posterior distribution $q(S)$ from a parameterized family of distribution over the latent variables with its own variational parameters $q(\mathbf{s}_i|\boldsymbol{\nu}_i)$. Within the parameterized family of distribution, we find the setting of parameters $\boldsymbol{\nu}_i$ such that it makes $q(\mathbf{s}_i)$ close to true posterior distribution $p(S|X)$. We measure the distance between approximate posterior distribution $q(S)$ and true posterior distribution $p(S|X)$ with Kullback-Leibler divergence:

$$KL\Big(q(S)||p(S|X)\Big) = E_q\left[\log \frac{q(S)}{p(S|X)}\right]$$
$$= \log p(X) - E_q[\log p(S, X)] + E_q[\log q(S)],$$

where

$$\mathcal{L}(q) = E_q[\log p(X|S)] + E_q[\log p(S)] - E_q[\log q(S)].$$

is the evidence lower bound (ELBO) that we use as a proxy for minimizing the KL distance between $q(S)$ and $p(S|X)$. Maximizing ELBO is equivalent to minimizing KL divergence.

In mean field variational inference, we assume that the variational family factorizes as

$$\begin{aligned} q(S) =& \prod_{i=1}^{N} q(\mathbf{s}_i) \\ q(\mathbf{s}_i) =& \mathcal{N}(\mathbf{s}_i|\mu_{\mathbf{s}_i}, \Sigma_{\mathbf{s}_i}) \end{aligned}$$

$$S = \begin{bmatrix} \mathbf{s}_1 & \dots & \mathbf{s}_t & \dots & \mathbf{s}_T \end{bmatrix} = \begin{bmatrix} \mathbf{s}_1^T \\ \vdots \\ \mathbf{s}_i^T \\ \vdots \\ \mathbf{s}_N^T \end{bmatrix}$$

S is the stacked shared randomness with size $N \times T$. We use the subscript i and t to distinguish $\mathbf{s}_i$, the i-th row of S, with $\mathbf{s}_t$, the t-th column of S. We use multivariate gaussian distribution as approximate distribution for $q(\mathbf{s}_i)$ due to the fact that $\mathbf{s}_i$ are generated by Gaussian processes.

Using mean field variational inference method, we derive the update equations for variational parameters $\mu_{\mathbf{s}_i}$ and $\Sigma_{\mathbf{s}_i}$ for the posterior distribution $q(\mathbf{s}_i)$:

$$\Sigma_{\mathbf{s}_i} = \left(\mathbf{K}_{\mathbf{s}_i}^{-1} + \sum_{m=1}^{M} \frac{\sum_{k=1}^{N}(W_m)_{ki}^2}{\rho_m^2} I\right)^{-1}$$

$$\begin{aligned}\mu_{\mathbf{s}_i} &= \Sigma_{\mathbf{s}_i}\left(\sum_{m=1}^{M}\frac{1}{\rho_m^2}\sum_{j=1}^{N}(W_m)_{ji}(\mathbf{x}'_m)_j - \big(\sum_{m=1}^{M}\sum_{j\neq i}^{N}\frac{\sum_{k=1}^{N}(W_m)_{ki}(W_m)_{kj}}{\rho_m^2}\big)E_{-\mathbf{s}_i}\left[\mathbf{s}_j\right]\right) \\ &= \sum_{m=1}^{M}\frac{1}{\rho_m^2}\Sigma_{\mathbf{s}_i}\left(\sum_{k=1}^{N}(W_m)_{ki}\left((\mathbf{x}'_m)_k - \sum_{j\neq i}^{N}(W_m)_{kj}E_{-\mathbf{s}_i}\left[\mathbf{s}_j\right]\right)\right).\end{aligned}$$

For hyperparameter estimation, we apply empirical Bayes method by updating hyperparameters with gradient-based optimization of ELBO.

4.5.3 Discussion and Conclusion

	SRM	GPSRM
Computation	$\mathcal{O}(vk^2)$	$\mathcal{O}(dk^2)$
Memory	$\mathcal{O}(dk + dk^2)$	$\mathcal{O}(kd + kd^2)$

Table 4.2: Comparison of computation and memory complexity.

In this section, we develop an extension to the shared response model with temporal dependency using Gaussian process, which we called GPSRM. Different from SRM's temporal independent assumption, we adopt a Gaussian process prior over the latent variable of SRM such that it will incorporate temporal structure. A variational inference algorithm for SRM is derived, which learns a posterior distribution over the latent variable. The computational bottleneck for GPSRM depends on the number of observations instead of numbers of voxels (Table. 4.2), which is suitable for dataset with large number of voxels but a small number of TRs. However, our empirical evaluation of this model suggests similar performance comparing to SRM. Even the time segment matching experiment, which we thought would benefit from this design,

demonstrates similar predictive performance to SRM. Our original hope was to model temporal structure with Gaussian process for learning a better shared feature space. However, there are a couple aspects that might prevent us from estimating a more informative shared feature space. First, we may not have a stationary latent time series structure that can be well modeled by GP. Second, we may not have enough data for explicitly modeling the temporal structure of the late variables. These future directions are compatible with the future directions in KSRM, and they are closely tied to the notion of using a nonlinear method for fMRI analysis.

Chapter 5

A Multi-view Convolutional Autoencoder

5.1 Introduction

Factor models operate on the principle of aggregating information across one or more dimensions of the data (space, time, subject). For example, SRM and Hyperalignment [60] aggregate information across space and subjects. Since aggregating across space (voxels) reduces anatomical spatial locality, these methods are usually applied to large pre-selected regions of interest (ROI), e.g., ventral temporal cortex [60] and posterior medial cortex [27]. Applying the models in this way can yield significant gains over prior methods in identifying informative responses in pre-selected regions [27]. However, these methods suffer from an important limitation: a lack of spatial locality. That is, all voxels within the selected region may contribute to the measure that is ultimately derived (e.g., a classification score). This limitation is at odds with a fundamental goal of neuroscience, which is to determine how local brain regions are associated with specific cognitive functions. For example, the ventral temporal cortex is known to contain a multitude of sub-areas, each with its specialized function

[49]. If all of these sub-areas enter an analysis together, overall classification scores may improve, but the ability to make inferences about the functional properties of individual sub-areas is lost.

Here we focus on the preservation of spatial locality during whole brain multi-subject data aggregation, with the aim of improving anatomical and functional interpretability of the analysis results. By preserving spatial locality we mean that information is only aggregated in a small region (e.g. a ball) about each voxel. A natural approach that can satisfy this constraint to combine factor models and searchlight based analysis [77, 41]. Searchlight analysis uses a small window of contiguous voxels around a known location to conduct a spatially local analysis. This analysis is performed at all locations in the volume. This generalizes an ROI approach to multiple (overlapping) spatially local "searchlights" across the brain. To handle m subjects, the analysis can be performed across m linked and co-centered searchlights, one per subject. This provides multi-subject, local data aggregation tailored to each searchlight [50]. In this chapter, we focus on this approach with the aim of making a connection between searchlight analysis and convolution neural networks. Other approaches that aim to ensure spatial locality are also possible. For example, a data-driven approach that learns "soft" boundaries of locally activated areas.

We explore the application of searchlights in two distinct ways: by combining the SRM with searchlights (S-SRM) (§3.7) and by using a multi-view convolutional autoencoder (CAE). A searchlight version of SRM is not conceptually new; similar idea has also been introduced in [50]. We bring it in as a fairer benchmark for the CAE than factor models without spatial locality constraints. To understand the relevance of a convolutional autoencoder we first note in §5.3 that a two layer fully connected autoencoder can replicate the performance of SRM on multi-subject fMRI data. But like the SRM, this autoencoder does not have spatial locality. We then argue that we can add spatial locality by transitioning from a fully connected to a

convolutional autoencoder. To see this consider the post-training application of an S-SRM analysis and that of a single layer convolutional neural network (CNN). In an S-SRM analysis, a fixed sized window is moved over the data and at each location we form k inner products between weight vectors (learned functional topographies in Chapter 2) and the windowed data. Similarly, in a convolution using k filters, a fixed size filter support is moved over the data and at each location k inner products of filter coefficient vectors and the windowed data are computed. In both cases, the results are recorded and indexed by the coordinates of the region center. Subsequent analysis is then based on the outputs produced in each case. While the above analogy shows a clear similarity, the two approaches also differ in important ways. First, in a S-SRM the weight vectors can depend on the searchlight index but in a CNN the filter weights are invariant with location. Thus the searchlight approach has a key advantage: it can vary data aggregation depending on anatomical location. Second, the SRM (and many other factor methods) impose a nonlinear geometric constraint on the weight vectors (e.g., orthonormality), whereas CNN filter weights are not directly geometrically unconstrained except perhaps in norm. Third, a CNN contains distributed nonlinear activation functions whereas in a factor model, data factorization is a global nonlinear operation. It is well known, however, that a fully connected neural network can make use of its distributed activation functions to approximate nonlinear functions [62, 14].

There are several previous applications of deep learning to fMRI data. We review this recent literature and draw connections with our work. For unsupervised feature extraction, the l_1 regularized restricted Boltzmann machine has demonstrated comparable performance with ICA while giving more localized features [103]. One dimensional temporal convolutional autoencoders have been applied on fMRI data in matrix form (voxel-by-time) in a temporal convolutional neural network framework [43]. Recent work on classifying neuroimaging data used semi-supervised linear

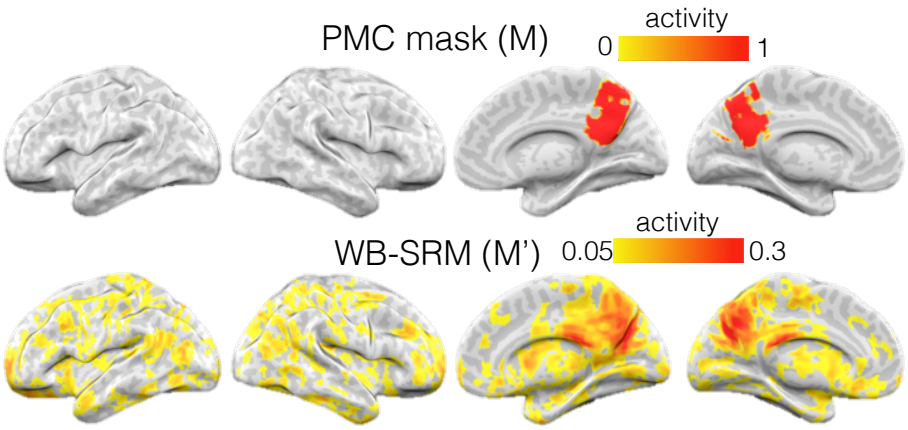


Figure 5.1: Illustration of the lack of spatial locality of whole brain SRM (WB-SRM) analysis. Figure from [29].

autoencoders to learn compressed representations of the neuroimaging data in the unsupervised stage [19]. Another classification paper uses deep neural networks to perform supervised learning, with the fMRI data as input and the corresponding class labels as output [76]. In this work, the intensities of voxels in each anatomical region of interest are averaged to help deal with the variability across subjects. These previous applications of deep learning to fMRI data do not explore co-activation across subjects nor the preservation of spatial locality in the aggregation of multi-subject data.

Our goal then is to design a multi-layer convolutional autoencoder for multi-subject, whole brain, spatially local, fMRI data aggregation. To do so, we create a network structure that matches the inherent multi-dataset nature of the problem and address some computational challenges arising from dealing with large-scale, multi-subject fMRI data. Our key contribution is to show that a suitably designed convolutional autoencoder can provide data aggregation that is competitive with methods based on whole brain searchlight analysis using latent factor methods. We also examine approaches to address the computational challenges of training a convolutional autoencoder using multi-subject fMRI data.

Prior Publications and Acknowledgment Parts of this chapter have been published in [29]. I thank Xia (Ivy) Zhu and Javier S. Turek for their tremendous help in optimizing the runtime performance of the neural networks.

5.2 Limitations of Current Methods

Let fMRI time-series data $X_i \in \mathbb{R}^{v_x \times v_y \times v_z \times d}, i = 1:m$, be collected from m subjects presented with an identical, temporally synchronized stimulus. The dimensions (x, y, z) of X_i are spatial coordinates in the brain, (v_x, v_y, v_z) are the number of voxels in the (x, y, z) dimensions, and d is the number of time samples in units of repetition time (TR). X_i can be regarded as a 4D tensor, but to afford wider accessibility we use standard multivariate notation. Our objective is to model, across the whole brain, the elicited response shared by the subjects while preserving its spatial locality. To do so we set out to identify local subject specific patterns that co-activate in time across subjects.

The SRM approach aims to achieve this goal by learning subject specific matrices $W_i \in \mathbb{R}^{v \times k}$, each with k orthonormal columns, and a shared response $S \in \mathbb{R}^{k \times d}$ to minimize the reconstruction error $\sum_{i=1}^{m} \frac{1}{m} \|X_i - W_i S\|_F^2$. Once learning is complete, one can project held out data X_i' for subject i into the shared response space by computing k inner products $S_i' = W_i^T X_i'$. One can also project this data into the voxel space of subject j by computing $W_j W_i^T X_i'$. The imposed orthonormality constraint plays a key role in achieving the performance reported in §3. If this constraint is removed, performance drops (§2.2.2). In addition, if spatial locality is desired, it must be externally imposed by restricting the SRM domain to a spatially local ROI. Applying the method across the whole brain forgoes spatial locality. We can demonstrate this using the *sherlock* dataset (see §5.5). After using the dataset to learn the SRM on the whole brain we obtain $W_i \in \mathbb{R}^{v \times k}$, $i = 1{:}m$ and a shared response

$S \in \mathbb{R}^{k \times d}$. We then create a synthetic brain map M in the voxel space of subject 1 taking value 1 in a post medial cortex (PMC) anatomical ROI and 0 elsewhere, and use the learned matrices W_1, W_2 to map M into the voxel space of subject 2: $M' = W_2 W_1^T M$. Preserving spatial locality requires that the support of M' is close to that of M. The result (Fig. 5.1) clearly shows that special locality is not preserved.

Our problem can also be conceived as multi-view learning problem and in this context, fully connected neural networks and autoencoders have proven useful [8, 127, 23]. It is possible to connect the SRM and a linear autoencoder by simply removing the constraint $W_i^T W_i = I_k$ and viewing the SRM objective as the reconstruction loss of a fully-connected linear, single hidden layer autoencoder (see Fig. 5.2). But dropping the above constraint reduces performance. In contrast, a nonlinear, multi-view autoencoder with two hidden layers (Fig. 5.3) can match the performance of SRM. However, like SRM, this autoencoder does not preserve spatial locality. Nevertheless, it suggests a novel approach to the fMRI data aggregation problem.

5.3 Fully-connected Autoencoder and SRM

We first show that SRM without the orthogonality constraint is equivalent to a tied-weights linear fully-connected multi-view autoencoder with one hidden layer as in Fig. 5.2. Since SRM does not keep track of spatial locality, the fMRI data is formulated by reshaping the 4D tensor response (v_x, v_y, v_z, d) into 2D response (v, d), where $v = v_x \times v_y \times v_z$. fMRI time-series data $X_i \in \mathbb{R}^{v \times d}$, $i = 1{:}m$, is collected for m subjects.

The core of SRM can be viewed as

$$\begin{aligned} &\min_{W_i, S} && \textstyle\sum_i \|X_i - W_i S\|_F^2 \\ &\text{s.t.} && W_i^T W_i = I_k, \end{aligned} \tag{5.1}$$

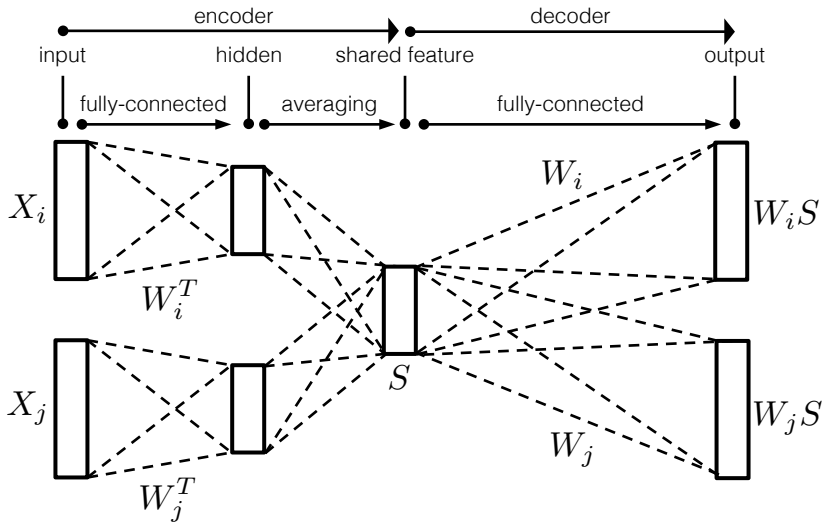


Figure 5.2: tied-weights linear fully-connected multi-view autoencoder. Figure from [29].

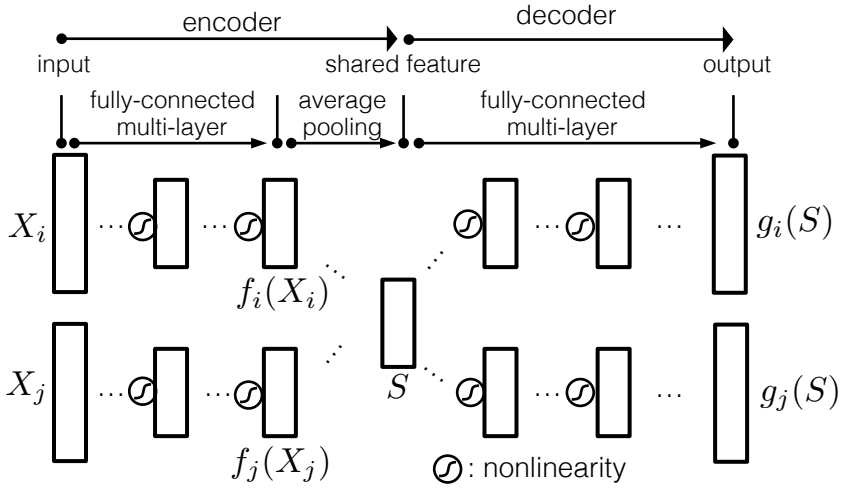


Figure 5.3: A nonlinear, fully connected autoencoder that can match the performance of SRM but also lacks spatial locality. Figure from [29].

where $\|\cdot\|_F$ denotes the Frobenius norm. Equation (5.1) can be solved iteratively by first initialize W_i, $i = 1{:}m$, and optimizing (5.1) with respect to S by setting $S = {}^1\!/_m \sum_i W_i^T X_i$. With S fixed, (5.1) becomes m separate Procrustes problems [47] of the form $\min \|X_i - W_i S\|_F^2$ with solution $W_i = \tilde{U}_i \tilde{V}_i^T$, where $\tilde{U}_i \tilde{\Sigma}_i \tilde{V}_i^T$ is SVD of $X_i S^T$ [61]. These two steps iterate until a stopping criterion is satisfied.

Using the fully-connected autoencoder, Fig. 5.2, input data X_i of subject i is transformed through subject specific fully-connected transformation W_i yielding $W_i^T X_i$ as hidden representation. The shared feature $S = {}^1/_m \sum_i W_i^T X_i$ is the average of each subjects' hidden representations similar to the formulation in SRM. From the shared feature, the decoder part of the network are tied-weight fully-connected transformations reconstructing the input from the shared feature S. The objective function of the network is loss between original data and reconstructed data $\min_{W_i} \sum_i \|X_i - W_i S\|_F^2$. In experiments similar to §3.4, this network performs similar to SRM without orthogonality but worse than SRM with orthogonality.

The linear fully-connected multi-view autoencoder, Fig. 5.2, can be generalized into a nonlinear network by introducing non-linearity. With the introduction of non-linearity, we design a multi-layer multi-view auto-encoder as in Fig. 5.3. The encoders for each subject's data X_i can be viewed as a subject specific nonlinear function $f_i(\cdot)$. Each subject's response in feature space is $f_i(X_i)$, and shared feature $S = {}^1/_m \sum_i f_i(X_i)$ is the average across subjects. From the shared feature S, the decoder network reconstructs the original input through a subject specific nonlinear functions $g_i(S)$. The whole network can be written as:

$$\min_{f_i, g_i} \quad \sum_i \|X_i - g_i(\tfrac{1}{m} \sum_j f_j(X_j))\|_F^2 + \lambda D_{\mathrm{KL}}(\rho \| \hat{\rho}). \tag{5.2}$$

The first term is the mean squared error between the reconstructed output $g_i(\frac{1}{m} \sum_j f_j(X_j))$ and each subject's data; the second term is the Kullback-Leibler (KL) divergence to a binomial distribution with parameter ρ [94]: $D_{\mathrm{KL}}(\rho \| \hat{\rho}) = \rho \log(\frac{\rho}{\hat{\rho}}) + (1 - \rho) \log(\frac{1-\rho}{1-\hat{\rho}})$ with ρ the desired sparsity and $\hat{\rho}$ the mean sparsity of the activations in the layer. This regularizes the network by sparsifying the shared feature maps S. Dropout is used to reduce overfitting [113]. We use the hyperbolic

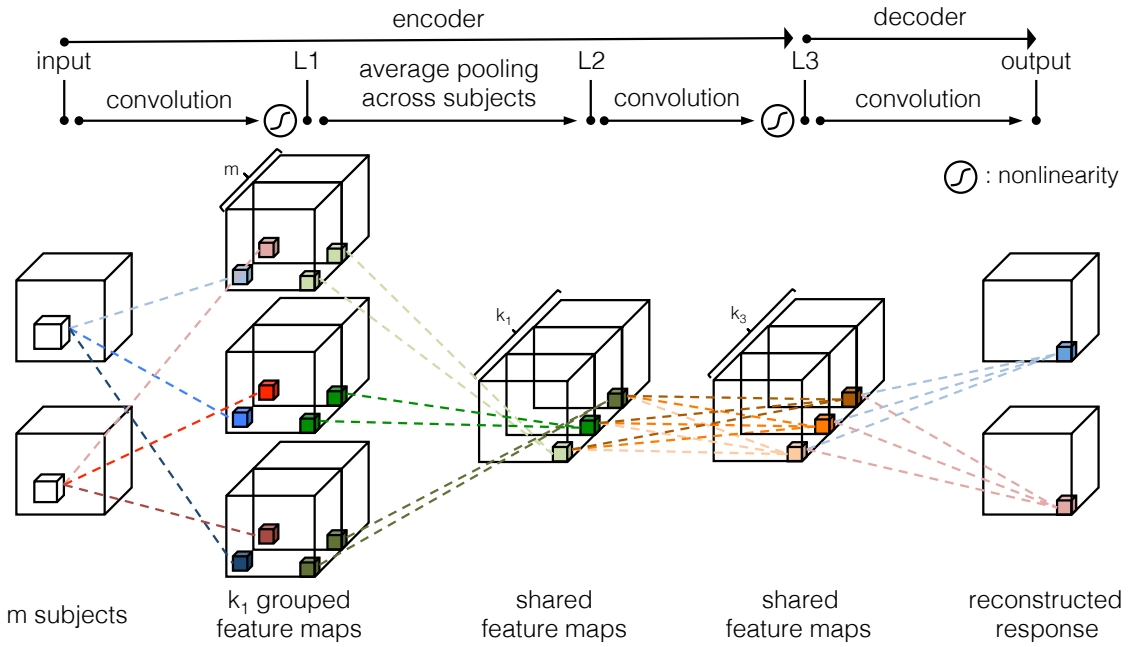


Figure 5.4: Proposed 4D convolutional autoencoder (CAE). Figure from [29].

tangent activation function since it yields shared feature maps with positive and negative values as in competing methods. We select the parameters ρ and λ using cross-validation and fix dropout on hidden layers to the typical value of 0.5 [113] and deactivate it on the input layer. In predictive experiments as in §3.4, this network leads to comparable performance as SRM.

5.4 A Multi-view Convolutional Autoencoder

Motivated by the desire for spatial locality, we now investigate a 4D convolutional autoencoder (Fig. 5.4) for multi-subject fMRI data aggregation. For simplicity, Fig. 5.4 shows only two subjects and we explain its operation in this context. The input data consists of 4D tensors $X_i, i = 1{:}m$ and the first layer is a 3D convolutional layer. To account for functional variability between subjects, this layer learns k_1 subject specific filters. However, filters with the same index are linked across the subjects. The output of the first layer is a set of mk_1 3D feature maps X_i^j; one per subject i and filter

index j. As shown in Fig. 5.4, these are subject-grouped for each linked filter index. The grouped feature maps specify the activity level across subjects of linked local filters at the locations across the brain. The second layer is average pooling across subjects. This identifies local patterns that co-activate across subjects (co-activating patterns). The result is k_1 shared feature maps. We do not expect activation of local spatial patterns alone to be informative of a shared response. Hence a second round of k_3 convolutions is performed over the k_1 activation patterns to identify local combinations of spatial activity patterns. This also introduces a second non-linearity into the network which is known to increase representational power [15]. The second convolutional layer computes k_3 1D convolutions resulting in k_3 3D feature maps. This design satisfies our goal of preserving spatial locality by aggregating the information across subjects from voxels within the filter support size. Each location in the final shared feature maps (L3) corresponds to brain searchlights linked across subjects. Finally, we use a single layer of convolutions to generate the reconstructed datasets $\hat{X}_i = h_{i,\theta}(X_1, \ldots, X_m)$, where θ is the model parameters. These represent the manifestation of the shared response in each subject's brain.

We train the convolutional autoencoder by minimizing the loss function

$$L(\theta; X) = \frac{1}{m} \sum_{i=1}^{m} \|X_i - h_{i,\theta}(X_1, \ldots, X_m)\|_F^2 + \lambda D_{\mathrm{KL}}(\rho \| \hat{\rho}). \tag{5.3}$$

The first term is the mean squared error between the reconstructed output $h_{i,\theta}(X_1, \ldots, X_m)$ and the subject's data; the second term is the Kullback-Leibler (KL) divergence to a binomial distribution with parameter ρ [94]: $D_{\mathrm{KL}}(\rho \| \hat{\rho}) = \rho \log(\frac{\rho}{\hat{\rho}}) + (1-\rho) \log(\frac{1-\rho}{1-\hat{\rho}})$ with ρ the desired sparsity and $\hat{\rho}$ the mean sparsity of the activation in the layer. This regularizes the network by sparsifying the k_3 shared feature maps in layer 3. We use the hyperbolic tangent activation function since the data is z-scored and it yields shared feature maps with positive and negative

Dataset	TRs (s/TR)	Voxel Region	# Voxels
audiobook [137]	449(2)	whole brain (WB) in MNI [88]	70273
sherlock-movie [25]	1976(2)	whole brain (WB) in MNI [88]	70273
		posterior medial cortex (PMC)[87]	813
sherlock-recall [25]	437∼1009(2)	whole brain (WB) in MNI [88]	70273
		posterior medial cortex (PMC)[87]	813

Table 5.1: fMRI datasets are shown in the left two columns, and the ROIs are shown in right two columns. We use 9 subjects from version of datasets that match the data in the corresponding publications.

values. The sparsity regularization is computed by scaling and shifting the hyperbolic tangent output to $[0, 1]$. Dropout is used to reduce overfitting [113]. We select the parameters ρ and λ using cross-validation and fix dropout on hidden layers to the typical value of 0.5 [113] and deactivate it on the input layer.

In our convolutional autoencoder, the number of model parameters is much smaller than the number of activations. Therefore, we adopt a data parallel method for distributed training that reduces communication overhead. We implement a distributed training framework for Theano [5] based on a synchronous Stochastic Gradient Descend (SGD) [69, 26] to handle the computational load for training the network. We select a synchronous method over asynchronous SGD because of the better convergence properties [26]. The synchronous SGD method has many processes running in parallel, each maintaining a copy of the entire model. Every SGD iteration, a mini batch is assigned to each process to compute a local gradient. Then, all these gradients are aggregated by a binomial reduction tree based collective operation. Eventually, the local models are updated using the aggregated gradient. In addition, we initialize all filters in the first layer with values from a random orthogonal matrix. We use RMSprop [116] to adaptively adjust the learning rate. For decay rate and smoothing value, we swept in the range $\{0.9, 0.99, 0.999\}$ and $\{10^{-4}, 10^{-6}, 10^{-8}\}$, respectively. The initial learning rate depends on the batch size and the number of nodes used.

5.5 Experiments and Results

The performance of S-SRM and the CAE was evaluated using two fMRI datasets: the *sherlock* dataset (§3.2.2) (with both *sherlock-movie* and *sherlock-recall*) and the *audiobook* dataset (§3.2.3). We use either whole brain data with 70273 voxels or posterior medial cortex (PMC) data with 813 voxels. Details are given in Table 5.1. The primary metrics are prediction accuracy, used as a proxy for relevant shared information, and spatial locality. The high accuracy regions suggest the presence of information relevant to the predictive tasks being conducted.

For the CAE we use a $5\times5\times5$ support region in first layer convolutions for full resolution fMRI data and $3\times3\times3$ regions for data down-sampled by 2. After training, held out data is mapped from the input layer to the shared response (bypassing across subject pooling) at the output of layer 3. S-SRM uses searchlights sized as above. Each searchlight contains v_s voxels and we use $k = 10$ features per searchlight. The training data for voxels in the m linked searchlights across subjects are used to learn an SRM and the learned subject-specific maps $W_i \in \mathbb{R}^{v_s \times k}$ are used to project held out data into the shared space for each searchlight. For both the CAE and S-SRM we then conduct time segment matching and brain map matching experiments using the data of a held out subject. The resulting accuracy of the matching tasks is assigned to the center voxel of the corresponding input region for CAE or the searchlight for S-SRM. This enables us to plot a local accuracy map across the whole brain. In comparisons with standard searchlight based analysis we use the same sized searchlights and for comparisons with single region SRM (whole brain or ROI) we use $k = 100$ features.

For training the CAE, we use the distributed synchronous SGD described in §5.4, applying MPI parallelism at the mini batch level. In addition, we tune Theano to make full use of OpenMP. Moreover, certain operations in NumPy and SciPy versions run serially and therefore we develop NumPy extension modules in C++ parallelized with OpenMP to speed these up. These optimizations yield up to 67× training

speedup on a single node comparing to the original Theano version, and an additional 7× speedup on an 8-node CPU cluster. Furthermore, we only load necessary data according to the mini batch to maintain a reduced memory footprint in each process.

We run our experiments on an 8-node cluster, interconnected by an Arista 10GE switch. Each node of the cluster has a motherboard with 2 Intel Xeon E5-2670 processors, both running at 2.6GHz, and with 256GB memory. The convolutional autoencoder is implemented in Python, with OpenMP for multi-threading, and mpi4py for multi-node parallelism. The software packages used in our experiments include Intel optimized Theano [5] (version rel-0.8.0rc1). We use the Anaconda distribution of Python with the following packages: Intel MKL 11.3.1, NumPy 1.10.4 and SciPy 0.17.0.

Experiment 1: Local region time segment matching.

We first use the *sherlock-movie* and *audiobook* datasets to replicate an experiment from §3.4. The experiment compares a standard searchlight analysis (SL) with S-SRM and the CAE. For each dataset, the movie data was split into halves, one-half was used for training the other for testing; then the roles were reversed and results averaged. The experiment tests if a 9 TR time segment from the testing data of a held-out subject can be located in the testing data of the subjects used in training. In the testing phase, we map subject's testing data from the input to the shared feature map (without conducting average pooling across subjects). A random 9 TR test segment from the testing half of the held-out subject's data is projected onto the shared space and we locate this segment in the averaged shared response of the other subject's testing data by maximizing Pearson correlation. Segments overlapping with the test segment are excluded from the matching process. We record average accuracy and standard error by two-fold cross-validation over the data halves and leave-one-out over subjects. Each dataset is in MNI space [88]. The accuracy maps

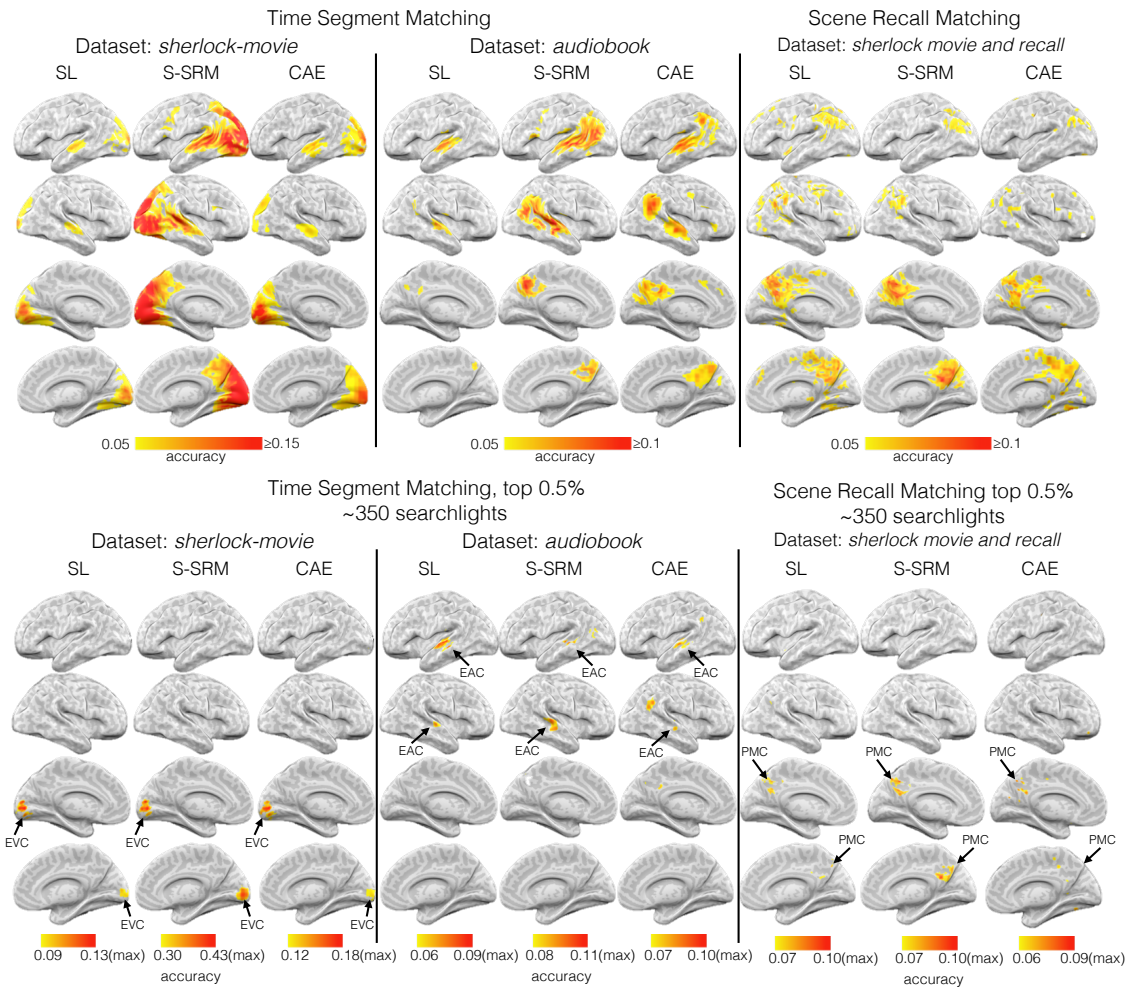


Figure 5.5: Top Left: Accuracy maps for Experiment 1 using *sherlock-movie* and *audiobook*; Top Right: Accuracy maps for Experiment 2 using *sherlock-movie* and *sherlock-recall*. Top figures are thresholded at corresponding scales for visualization clarity purpose. Please refer to bottom row figures for the high end of the range. Bottom Left: Accuracy maps for top 0.5% searchlights for Exp. 1; Bottom Right: Accuracy maps of top 0.5% searchlights for Exp. 2. Early Visual Cortex (EVC), Early Auditory Cortex (EAC), Posterior Medial Cortex (PMC). Figures from [29].

are shown on the left of Fig. 5.5. Accuracies below 0.05 were set to zero. Since each searchlight contains only a small local view, its predictive performance is expected to be low. The experiment was also conducted using a univariate voxel test but no voxel scored above 0.05. Assuming independence, chance accuracy is 0.0044 for the *audiobook* dataset and 0.001 for *sherlock-movie* dataset.

Experiment 2: Scene recall matching.

We now use *sherlock-movie* and *sherlock-recall* to compare standard SL analysis, S-SRM and CAE analyses on a more challenging task. We label each TR of the *sherlock-recall* data with the corresponding scene based on the subject's verbal description. The TRs captured during a subject's recall of the same scene are averaged. Our goal is to test if subjects have a similar brain activation pattern when retrieving the memory of the same scene. To do so we attempt to classify the scene of the recall responses of a left out subject . The whole movie is used to train S-SRM and the CAE. The effectiveness of the learned shared response is then tested using data from a held out subject. After projecting the *sherlock-recall* data to the shared space, an SVM classifier is trained and the average classifier accuracy and standard error are recorded by leave-one-out across subject testing. The results are shown as the accuracy plots on the right in Fig. 5.5. Assuming independence, chance accuracy is 0.02.

Experiment 3: Whole brain time segment and scene recall matching.

In this experiment, we investigate how well we can perform time segment matching and scene recall matching using a classifier that combines locally learned information across the whole brain. This experiment compares five approaches: whole brain voxel analysis (VX), whole brain SRM (WB-SRM) with $k = 100$ features, standard SL, S-SRM, and CAE. The experiment procedure is similar to Exp. 1 and Exp. 2, however, instead of doing classification in each local region, we classify using the results of all the local analyses across the whole brain. Whole brain voxel analysis (VX) is done by directly calculating time segment matching on whole brain voxel data without any model. WB-SRM is done by applying SRM ($k = 100$) on whole brain data.

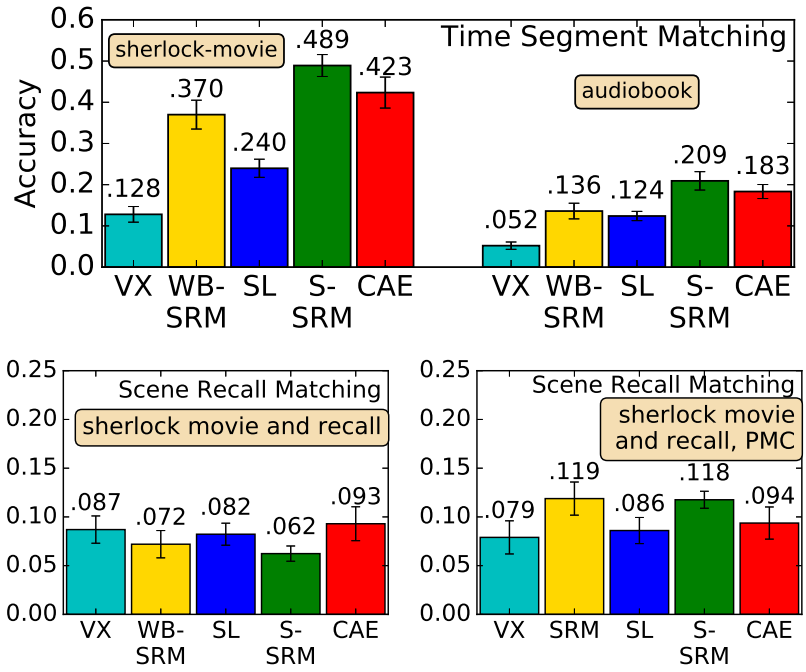


Figure 5.6: Prediction accuracies for Experiment 3. Top: Comparison of 9 TRs time segment matching on two datasets. Bottom Left: Comparison of movie scene recall classification on *sherlock*. Bottom Right: Comparison of movie scene recall classification on *sherlock* in PMC. Error bars: ±1 stand. error. Figures from [29].

Standard searchlight analysis (SL), S-SRM and CAE are local methods applied as before directly to whole brain data. By aggregating the local information from all local regions, we expect higher predictive power. The results are shown in the top plot of Fig. 5.6 for time segment matching. From time segment matching results, we observe significantly better predictive performance when combining information from both searchlight spatial locality and functional shared feature, e.g. S-SRM and CAE. On the other hand, without spatial locality information, WB-SRM shows slightly worse performance. Without functional shared feature information, SL also shows worse performance. Without information from both spatial locality and functional shared feature, VX shows worst performance.

Whole brain scene recall results are shown in the bottom left plot of Fig. 5.6. Motivated by these results, we also conduct the same scene recall experiment in the PMC ROI. PMC is known to be informative for recall classification [25]. These

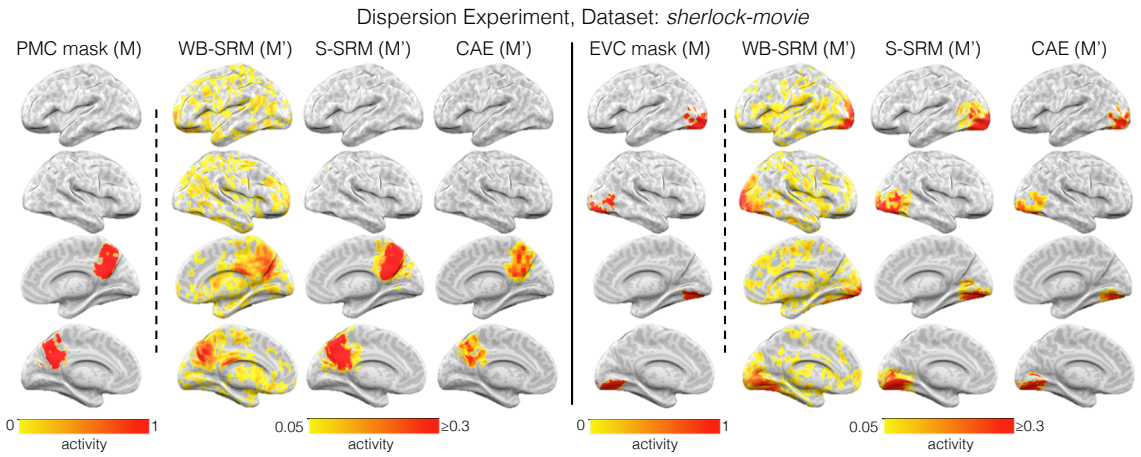


Figure 5.7: Experiment 4. A dispersion comparison between S-SRM and CAE using two anatomical ROI masks: Posterior Medial Cortex (PMC) and Early Visual Cortex (EVC). Figures from [29].

results are shown in the bottom right plot of Fig. 5.6. WB-SRM and S-SRM both show improvement in accuracy while using only voxels in PMC. This suggests that if the distribution of information in the brain is already known, it is better to use SRM with only voxels from the ROI to achieve better performance. However, if the distribution of information is unknown, using CAE and SL provide robust predictive performance while uninformative voxels are included.

Experiment 4: Dispersion

We now examine how well S-SRM and the CAE address the issue of spatial locality. We conduct the same experiment described in §5.2 and shown in Fig. 5.1. We use two ROI regions to compare the spatial dispersion of S-SRM and CAE (Fig. 5.7). As expected, S-SRM and the CAE have much less spatial dispersion than WB-SRM. The S-SRM exhibits slightly greater dispersion than the CAE.

5.6 Discussion and Conclusion

Our objective is to accomplish whole brain, multi-subject, fMRI data aggregation while preserving the spatial locality of information. The dispersion experiment (Fig. 5.7) indicates that both S-SRM and the CAE preserve spatial locality. The key remaining issue is whether aggregation of fMRI responses using these methods better distinguishes local and global cognitive states. To check this, we use the accuracy maps as a proxy measure of the effectiveness of the information aggregation.

The results of the time segment matching experiments indicate that both S-SRM and CAE enable improved matching of temporal segments over standard searchlight analysis (locally and globally) and marginally over WB-SRM (globally) (Fig. 5.5 (top left) and Fig. 5.6 (left)). For the *sherlock-movie* both S-SRM and CAE result in regions of the highest accuracy in the early visual cortex (EVC), which accords with neuroscience expectations for having early sensory areas driven in a specific and predictable way by the stimulus. On the *sherlock-movie* dataset the peak predictive performance averaged across the top 0.5% ($\approx$ 350 voxel locations) of the local regions is 0.11 for SL, 0.35 for S-SRM, and 0.14 for the CAE. The averaged peak accuracy of the local regions of S-SRM and CAE clearly outperform those of standard SL analysis. S-SRM has done the best job of local aggregation of information areas with high peak accuracy in EVC. The CAE is second in rank with lower peak accuracy but nevertheless good coverage of relevant brain areas. Its peak accuracy is also in EVC. Our claim is also supported by the results of whole brain classification (Fig. 5.6) where S-SRM and CAE attain the highest classification accuracies; distinctly above standard searchlight analysis. For *audiobook*, both S-SRM and CAE have comparable spatial performance with highest accuracy in early audio cortex (EAC). The peak predictive performance averaged across the top 0.5% ($\approx$ 350 voxel locations) of the local regions is 0.07 for SL, 0.10 for S-SRM, and 0.08 for the CAE. While the averaged peak accuracy of local regions of S-SRM and CAE are slightly higher than SL analysis, the

combined whole brain predictive accuracy (Fig. 5.6(left)) for both S-SRM and CAE are about twice as large as the best local region (Fig. 5.5 (bottom)).

It is particularly interesting that whole brain SRM (WB-SRM) does not perform as well as either of the local methods (S-SRM, CAE) when classifying temporal segments using whole brain data. This suggests that for cognitive state classification, it is better to perform a local spatial analysis first, then combine the results of the local analyses to perform a global prediction of the cognitive state.

The scene recall matching experiment provides a challenging task for all methods. We observe no improvement in whole brain classification over the best local prediction accuracies. This suggests that the relevant information is highly spatially localized. All three local analysis methods indicate that it is localized in the PMC ROI consistent with the finding in [25]. A follow-up experiment based only on PMC (Fig. 5.6 (right)) shows that the performance of the standard searchlight method and the CAE is the same when applied to the whole brain and when applied on PMC. On the other hand, the performance of SRM and S-SRM improves when restricted to PMC. This suggests that with prior knowledge of informative local regions, it's best to use SRM and S-SRM directly in the ROI.

A key distinction between scene recall matching and time segment matching is that the scene recall test probes representations at a higher level of stimulus processing. It is known that neural representations become more abstract at higher and higher levels in the processing stream (e.g., as one moves from early sensory areas up to areas like PMC) [57]. Responses in higher level areas are generally less similar across subjects, compared to early sensory regions, likely due to their intrinsically more complex relationship to the stimulus; this property is observed in the identification of low-level sensory areas EVC (for *sherlock-movie*) and EAC (for *audiobook*) as some of the most informative voxels in the time segment matching test (Fig. 5.5). When two data types are not matched regarding sensory input, these low-level areas do not

match; instead, we find that PMC carries the most strongly shared information (the scene recall matching test). PMC is at the highest level of the stimulus processing stream, and as our experiments have displayed, it has the interesting property of exhibiting similar response patterns for two scenes with similar content irrespective of the type of sensory input (movie vs. spoken recall) [25]. Successful decoding of cognitive state using local information in the brain helps determine a local brain region's specific cognitive function, and also demonstrates what kind of information is present and the information's distribution across the brain. S-SRM and CAE have shown increased sensitivity for both local and global investigation.

In summary, we have investigated and compared two ways of preserving spatial locality in multi-subject fMRI data aggregation: searchlight SRM and a convolutional autoencoder. Both approaches show improved results over standard competing methods. To our knowledge, the application of a convolutional autoencoder to this task is novel and moves away from factor model approaches which appear to be hitting a performance ceiling. With further refinement, a well-trained convolution autoencoder may lead to a more powerful means of accomplishing the fMRI data aggregation task.

Chapter 6

Beyond Multiclass Classification

6.1 Introduction

Scientific discovery is based on the scientific method of hypothesis testing. The prediction and classification approach that we adopt can also be viewed as a form of hypothesis testing. We have demonstrated this using ROI and searchlight analysis in various experiments of Chapter 3. Those experiments have answered lots of ***what*** and ***where*** questions. For example, where can we find what image stimulus category information in the brain; where does the brain exhibit different responses evoked by different interpretations of the same story, etc. On the contrary, ***why*** and ***how*** questions are harder to handle with fMRI data. For example, why does image category information show up in a particular region, how is this brain activation pattern formed, why is a particular brain activation pattern formed, etc. A primary reason for this limitation is the spatial and temporal resolution of fMRI. Even though fMRI is currently the best non-invasive neuroimaging technology, its spatial and temporal resolution is still insufficient to understand brain activation patterns in a bottom-up approach.

Encoding and decoding [93, 104] are two operations that indicate how the brain represents information. Encoding predicts brain activity given a stimulus, while decoding predicts stimulus given a brain activity. For most experiments in Chapter 3, we project data from voxel space into shared feature space and then conduct cross-subject stimulus prediction. This falls into the decoding paradigm. In a multi-view learning framework, by treating stimulus features as a view, we bridge between the stimulus feature space and fMRI voxel space. This approach makes encoding and decoding two sides of the same coin. Encoding can be viewed as transforming data from stimulus features to voxel space and decoding can be viewed as transforming data from voxel space to stimulus feature space.

In this chapter, we go beyond stimulus prediction as in most experiments of Chapter 3. The stimulus prediction problem requires training a classifier that takes either raw fMRI data or fMRI features as input and then outputs a stimulus label. In this setting, there are only a limited number of classes/labels we can predict. Furthermore, it requires training data from all possible labels to do a good job in prediction.

Here we test the possibility of learning linear transformations between fMRI shared feature space and word semantic space. The goal is to match the semantic representations via a linear transformation. Relating the fMRI data to the semantics of the stimulus (e.g. text, movie, etc) helps us understand the neural representation of semantic meaning. Furthermore, by successfully bridging word semantic space and shared feature space, we can transform data between stimulus features and brain activation patterns. This opens up the potential to go beyond answering the ***what*** question, and bring us closer to answering the ***how*** question, e.g., how is a particular stimulus feature represented in the brain.

There has been previous work trying to connect brain response and stimulus, such as words [92], pictures[46, 70], videos [96], stories [130, 65], mental images [115], etc. We review several related works in the domain of text. In [92], subjects view images

of concrete nouns (e.g. horse, dog) while fMRI data is collected. The authors try to predict fMRI voxel activity using 25-dimensional word features using a learned linear basis to map from word vectors to fMRI activation. Then a binary classification is conducted on held-out data. In [131], the authors extend the same approach from subjects reading individual words to subjects reading word sequences using a chapter of Harry Potter for predicting voxel activity. In [100], the authors start with Wikipedia articles to create a topic model [17]. With the topic model, the authors use the text representation as an approximation of the mental representation of the concept. In combination with the same dataset as in [92], the authors generate word clouds based on an fMRI response, using the generative properties of the topic model [100]. In [66], the authors study fMRI responses to a natural movie stimulus, represented with a feature space of 1705 distinct nouns and verbs. A subsequent study [65] analyzes fMRI responses to audio stories to derive a semantic word map for the voxels of the brain. In [101], the authors collect fMRI data from people who are shown words and short sentences paired with images of objects. With fMRI response evoked by this stimulus, the authors develop a model for learning a subject-independent mental representation of concepts and reconstructing generic mental representations from fMRI data.

The contribution of this chapter is twofold. First, we show that multi-subject fMRI data aggregating using SRM leads to improved decoding performance between fMRI and the text annotation. Second, we compare different ROIs on matching between fMRI and text annotation.

Prior Publications and Acknowledgment Parts of this chapter have been published in the [126]. This research is a collaborative effort. My primary contribution is the idea formulation, design of algorithms, and the design of the experiments. For other parts of the work, I'll briefly describe them in this chapter for continuity

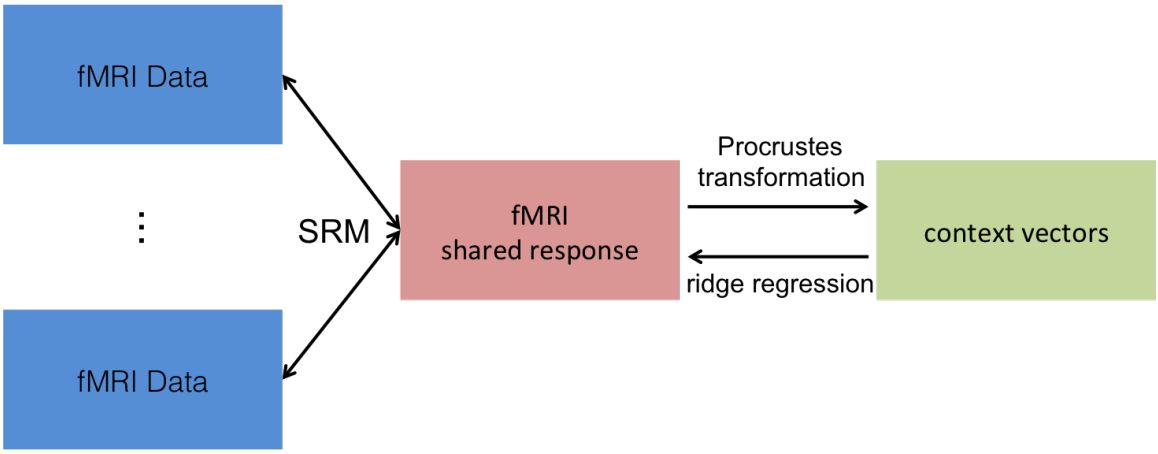


Figure 6.1: Illustration for experiment setup.

and self-containment purposes. The reader is referred to the above paper for further details. I thank Kiran Vodrahalli for his permission to use figures from [126].

6.2 Bridging Semantic Space and fMRI Space

We study the *sherlock* dataset (§3.2.2) with sub-second-resolution English text scene annotations. For each time point, we have the corresponding fMRI data and a textual description of the content in the movie. To bridge between the text annotation and the fMRI response, three key steps are used. First, we construct and aggregate semantic vectors from the text, particularly sequences of words. Second, we construct and aggregate fMRI shared features across subjects. Third, we bridge between the word semantic space and the fMRI shared feature space.

Learning semantic vectors We use natural language processing methods to represent words and sentences in a vector space. A common approach is to use word co-occurrence from a large corpus as the primary source of information to create word semantic vectors [99]. Then sentence semantic vectors can be estimated by averaging word semantic vectors. Recent approaches use neural network to estimate word and sentence semantic vectors by maximizing likelihood of a language model using neural

network [16, 79, 75]. In this work, we use the method proposed in [11] for estimating sentence semantic vectors. The method introduced in [11] constructs sentence semantic vectors using weighted average of word semantic vectors generated from one of the popular methods [79, 75, 132] with common component removal [11].

Identifying Shared Response We use SRM to identify a shared feature space among multi-subject fMRI dataset. We set dimension of the feature space k to 20. This number was selected because a low-rank SVD with 20 dimensions captures 90% of the variance of the original fMRI data. We then project data to the shared feature space using the learned basis W_i to estimate shared features. The estimated shared features can be written as $\frac{1}{N}\sum_{i=1}^{N} W_i^T X_i^{\text{test}}$. Four different algorithms are used for comparison, including simple averaging of the different responses $\frac{1}{N}\sum_{i=1}^{N} X_i^{\text{test}}$, PCA, SRM, and SR-ICA (§4.3).

Bridging fMRI Shared Response and Semantic Vectors To bridge between fMRI shared features and semantic vectors, we use linear regression. We use orthogonal regularization for mapping from the fMRI shared feature space to the semantic vector space. We use ridge regularization for mapping from the semantic vector space to the fMRI shared feature space. Let $X \in \mathbb{R}^{v \times d}$ be fMRI shared feature and $Y \in \mathbb{R}^{100 \times d}$ be the semantic vectors, where v is the number of voxels, d is the number of observations, and 100 is the selected dimensionality for semantic vector space. The first formulation with orthogonal regularization can be written as

$$\begin{aligned} &\min \|Y - \Omega_{XY} X\|_2^2 \\ &\text{s.t. } \Omega_{XY}^T \Omega_{XY} = I_k, \end{aligned} \tag{6.1}$$

where $\Omega_{XY} \in \mathbb{R}^{100 \times v}$ as a transformation from X to Y. In other words, this formulation decodes fMRI shared features into the semantic space.

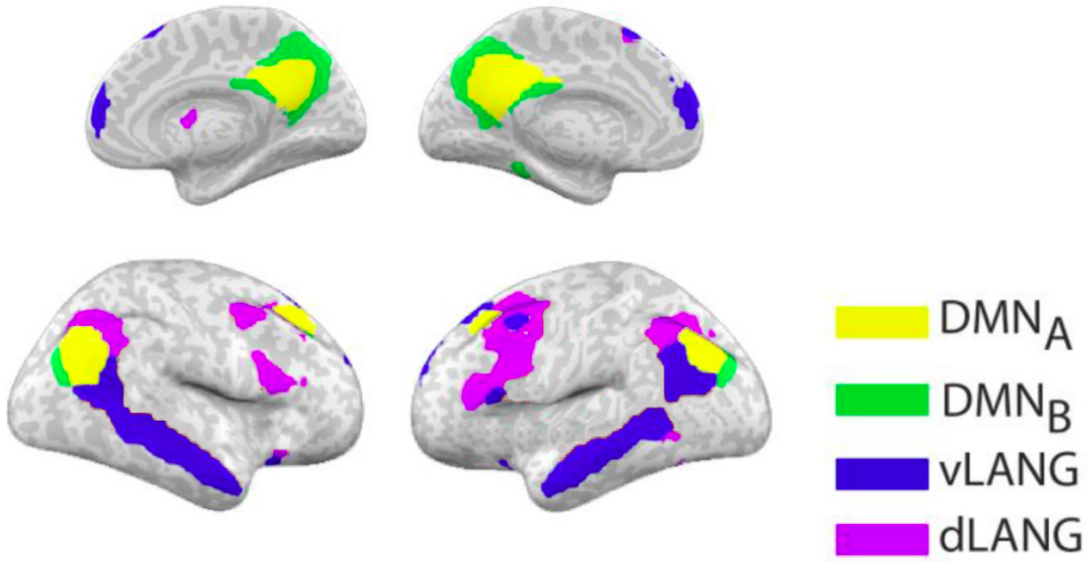


Figure 6.2: Visualization of the default mode network (DMN) and ventral/dorsal language area (vLANG/dLANG) ROIs. Figures from [126].

For encoding semantic vectors into the fMRI shared feature space, from Y to X, we adopt ridge regression,

$$\min \|X^T - Y^T \Omega_{YX}\|_2^2 + \|\Omega_{YX}\|_2^2. \tag{6.2}$$

where $\Omega_{YX} \in \mathbb{R}^{100 \times v}$ as a transformation from Y to X. Combing Ω_{YX} with the learned bases W_i in SRM, we can map a semantic vector into the brain to construct a corresponding brain activity.

6.3 Experiments

We use the *sherlock* dataset (§3.2.2) with text annotation of the scenes [25]. Voxels from multiple ROIs (see Fig. 6.2) are selected for comparison and evaluation. We conduct two experiments, scene classification, and scene ranking, for evaluating the effectiveness of our approach to bridging the semantic space and fMRI space. We

first equally divide the dataset into 50 chunks in time. Half of the chunks are used as training data, and the other half are used as testing data. We learn the mapping from fMRI to text and from text to fMRI using the training data on individual TRs instead of 25 chunks. To predict text from fMRI, for each time chunk $i \in [1, 25]$ in fMRI space, we predict chunk i in semantic space using the learned transformation. We then calculate the Pearson correlation of the predicted chunk i with each of the true chunks $j \in [1, 25]$, and rank the chunk indexes by correlation. There are two different performance metrics we adopt in this case. In scene classification, a classification is correct if the true chunk index is ranked among the top 5 chunks produced by this sorting. Hence the chance rate is 20%, and closer to 1 is better. In scene ranking, we report $1 - \frac{\text{average rank of the true index}}{25}$, which has 50% chance rate and also closer to 1 is better. The main reason for reporting two different metrics is to provide a better understanding of the ranking distribution.

In Fig. 6.3, we show the accuracy for our four experiments over different ROIs: they are fMRI to text on classification, fMRI to text on ranking, text to fMRI on classification, text to fMRI on ranking. Linear maps estimated by equation (6.1) and equation (6.2) are used for fMRI to text and text to fMRI. For fMRI to text, equation (6.1) performs 1.25x better than equation (6.2) in testing accuracy. However, for text to fMRI, equation (6.2) performs 1.2x better than equation (6.1) in testing accuracy.

We observe 72% testing accuracy for the scene classification task for fMRI to text and mid-90%s testing accuracy for the scene ranking tasks. The ROI DMN outperforms the others, which is consistent with the results in [106, 111] and other works demonstrating that the DMN is critical to narrative processing. The highest accuracy achieved is 96% accuracy over 20% chance for the scene classification task for mapping fMRI to text. We consistently achieve $> 80\%$ accuracy in all ROIs for both measures.

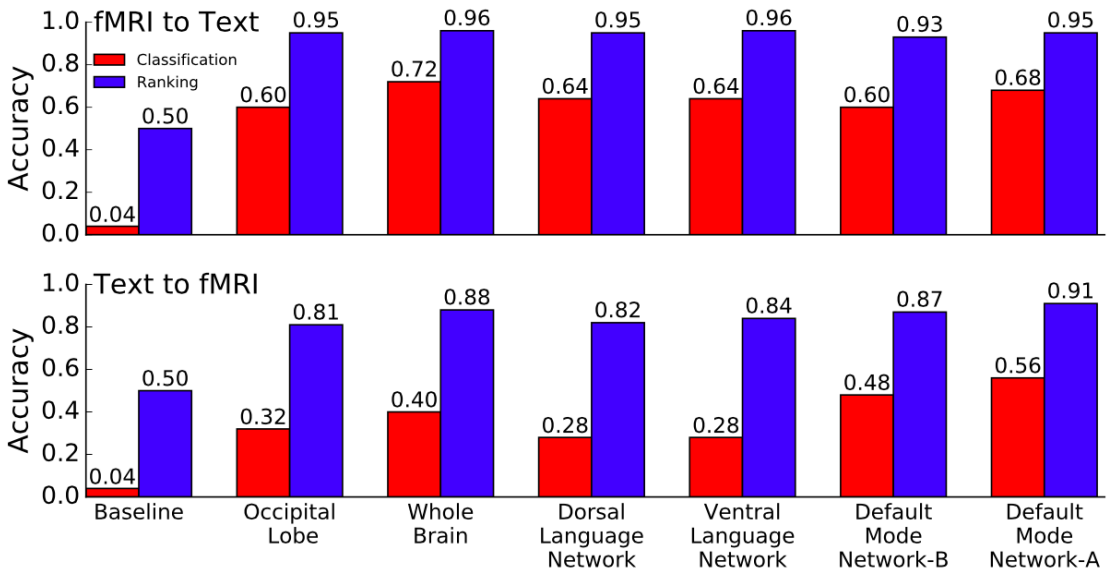


Figure 6.3: Bidirectional accuracy score for each brain region of interest for both scene classification and ranking (std. err. over different average subsets ≤ 0.01). Figures from [126]

On the other hand, for text to fMRI experiments, we observe worse performance than the fMRI to text experiments. The best scene classification accuracy performance is 56% for the DMN-A region, and the other top performing regions get accuracy in the mid-to-high 40% accuracy. For the ranking task, performance ranges from 80% to 90%, which is again slightly worse than the fMRI to text ranking experiment.

6.4 Discussion and Conclusion

Treating fMRI movie watching data and text annotation of the movie as different views of the same underlying representation, we have employed multi-view techniques to bridge between fMRI data and text. Two levels of multi-view learning are used here. In the first level, we treat fMRI data from each subject as a view and aggregate fMRI data across subjects using SRM. The second level of multi-view representation learning is used between fMRI shared features and text. Text annotation of the movie is first transformed into context vectors using word embedding techniques. With the

fMRI shared feature space and context vectors, we use Procrustes transformation and ridge regression to learn mappings from fMRI to text and from text to fMRI, respectively. A classification and a ranking experiments are conducted to evaluate the effectiveness of using linear transformations to bridge between the two spaces. The experimental results demonstrate high accuracy in matching generated fMRI data with true fMRI data as well as generated context vectors with true context vectors, suggesting the potential of our methods to bridge between the two spaces.

In this work, the experiments were done in two phases. Moving towards an end-to-end approach is an interesting and plausible future direction. An end-to-end approach is to use a single model that directly maps raw fMRI to text and from text to fMRI. A possible approach to achieve this is through a neural network formulation bridging between fMRI and text. Motivated by image caption generation [135, 125] from computer vision domain, with an architecture like this, a possible experiment can be something like brain caption generation. However, instead of using a natural image, we use fMRI data as input, and the output is text describing aspects of interest in the fMRI data.

Chapter 7

Conclusion

The application of multi-view representation learning to neuroimaging data is at an early stage. We hope this thesis has not only shed light on using multi-view representation learning with functional neuroimaging data but will also lead to further understanding and progress in both the machine learning and neuroscience communities.

In this thesis, we first propose a generic multi-view representation framework based on a factor model. A particular instance from the framework, called the shared response model, was fully developed in a probabilistic setting. The SRM learns a low dimensional latent representation across views capturing within-view variance and pairwise covariance. Furthermore, by assuming data from multiple views as different realizations of the same underlying source, the model can overcome a mismatch in view specific representations. This allows the model to learn a common low-dimensional shared feature space regardless of the difference in dimensionalities between view specific observations. Lastly, this framework can be easily extended for different purposes in fMRI data analysis. For example, we have extended the framework to train on both labeled and unlabeled datasets in a semi-supervised way

and have modified the objective function to make it more suitable to the scientific question being worked on.

So, how can this help us learn more about the brain? First, the ability to effectively aggregate data over multiple subjects opens up the opportunity to better utilize much larger datasets for scientific analysis. The key advantages of this are increased statistical sensitivity via the usage of a significantly larger dataset and generalizability of the results.

Second, by applying the framework in a spatially confined region, we can learn more about the distribution of information in the brain. With this in mind, fMRI is very good at answering ***what*** and ***where*** questions. We have demonstrated this using ROI and searchlight analysis in various experiments of Chapter 3, Chapter 4, and Chapter 5. On the contrary, ***why*** and ***how*** questions are harder to handle with fMRI data. However, in Chapter 6, we used multi-view representation learning model to bridge between stimulus features and fMRI response. This opens up new ways to analyze data and brings us closer to answer these types of questions.

Third, the notion of multi-view learning also opens up new possibilities for analyzing neuroimaging data. Through the lens of multi-view representation learning, different scientific problems can be viewed under a uniform framework. For example, fMRI multi-subject functional aggregation can be viewed as multi-view learning problems by treating each subject as a view. Encoding and decoding problems can be viewed as a multi-view learning problems by treating fMRI and stimulus features as different views, encoding is mapping stimulus feature into voxel space, and decoding is the other way around. The encoding and decoding approaches open up the potential to go beyond only answering ***what*** questions, bringing us closer to answering ***why*** and ***how*** questions, e.g. how is a particular stimulus feature represented in the brain. This is partially demonstrated in Chapter 6.

Lastly, as a member of the computational science community, we have the privilege to conduct research with data and computation power which can be easily replicated. I do believe that this kind privilege also comes with the corresponding responsibility of making publicly available reproducible research. I have been emphasizing reproducible research and trying to make my research results practically usable by sharing my code on github[1] and using publicly available datasets in my experiments. Furthermore, I have been contributing my research results to open-source software, e.g. brainIAK[2], PyMVPA[3] [53, 54], two popular open-source toolboxes designed for neuroimaging research.

[1] www.github.com/cameronphchen/
[2] brainiak.org
[3] www.github.com/PyMVPA

Appendix A

Prior Presentations and Publications

A.1 Prior Presentations

1. A semi-supervised method for multi-subject fMRI functional alignment Javier S. Turek, Theodore L. Willke, Po-Hsuan Chen, Peter J. Ramadge ICASSP, 2017

2. Enabling Factor Analysis on Thousand-Subject Neuroimaging Datasets, Michael J. Anderson, Mihai Capota, Javier S. Turek, Xia Zhu, Theodore L. Willke, Yida Wang, Po-Hsuan Chen, Jeremy R. Manning, Peter J. Ramadge, Kenneth A. Norman, IEEE Big Data, 2017

3. A Convolutional Autoencoder for Multi-Subject fMRI Data Aggregation, Po-Hsuan Chen, Xia Zhu, Hejia Zhang, Javier S. Turek, Janice Chen, Theodore L. Willke, Uri Hasson, Peter J. Ramadge, Representation Learning in Artificial and Biological Neural Networks workshop, NIPS, 2016

4. A Searchlight Factor Model Approach for Locating Shared Information in Multi-Subject fMRI Analysis, Hejia Zhang, Po-Hsuan Chen, Janice Chen, Xia Zhu,

Javier S Turek, Theodore L Willke, Uri Hasson, Peter J Ramadge, Brains and Bits: Neuroscience Meets Machine Learning, NIPS 2016.

5. Mapping Between Natural Movie fMRI Responses and Word-Sequence Representations, Kiran Vodrahalli, Po-Hsuan Chen, Yingyu Liang, Janice Chen, Esther Yong, Christopher Honey, Peter Ramadge, Kenneth Norman and Sanjeev Arora, Representation Learning in Artificial and Biological Neural Networks workshop, NIPS, 2016.

6. A Semantic Shared Response Model, Kiran Vodrahalli, Po-Hsuan Chen, Janice Chen, Esther Yong, Christopher Honey, Kenneth Norman, Peter Ramadge and Sanjeev Arora, In Workshop on Multi-View Representation Learning, ICML, 2016

7. Kernelized Shared Response Model, Po-Hsuan Chen, Peter Ramadge, In 10th Annual Machine Learning Conference, NYAS, 2016.

8. A Reduced-Dimension fMRI Shared Response Model, Po-Hsuan Chen, Janice Chen, Yaara Yeshurun-Dishon, Uri Hasson, James Haxby, Peter Ramadge, Advances in Neural Information Processing Systems (NIPS), 2015.

9. A probabilistic latent factor approach for multi-subject fMRI data modeling, Po-Hsuan Chen, Peter Ramadge, Society for Neuroscience Abstracts, 2015.

10. Probabilistic hyperalignment, Po-Hsuan Chen, Peter J. Ramadge, In 9th Annual Machine Learning Conference, NYAS, 2015.

11. Probabilistic hyperalignment, Po-Hsuan Chen, Peter Ramadge, Workshop on Machine Learning and Interpretation in Neuroimaging (MLINI), NIPS, 2014

12. Joint SVD-Hyperalignment for multi-subject FMRI data alignment, Po-Hsuan Chen, J. Swaroop Guntupalli, James V. Haxby, and Peter J. Ramadge, IEEE Machine Learning for Signal Processing (MLSP), 2014

13. Joint SVD as warm start for hyperalignment, Po-Hsuan Chen, J. Swaroop Guntupalli, James V. Haxby, and Peter J. Ramadge, In 8th Annual Machine Learning Conference, NYAS, 2014.

A.2 Prior Publications

1. Po-Hsuan Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A reduced-dimension fMRI shared response model. In Advances in Neural Information Processing Systems, pages 460468, 2015.

2. Po-Hsuan Chen, J Swaroop Guntupalli, James V Haxby, and Peter J Ramadge. Joint SVD-Hyperalignment for multi-subject fMRI data alignment. In IEEE International Workshop on Machine Learning for Signal Processing (MLSP), pages 16, 2014.

3. Po-Hsuan Chen, Xia Zhu, Hejia Zhang, Javier S Turek, Janice Chen, Theodore L Willke, Uri Hasson, and Peter J Ramadge. A convolutional autoencoder for multi-subject fMRI data aggregation. arXiv preprint arXiv:1608.04846, 2016.

4. Javier S. Turek, Theodore L. Willke, Po-Hsuan Chen, and Peter J. Ramadge. A semi-supervised method for multi-subject fMRI functional alignment. In IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2017.

5. Kiran Vodrahalli, Po-Hsuan Chen, Yingyu Liang, Janice Chen, Esther Yong, Christopher Honey, Peter Ramadge, Ken Norman, and Sanjeev Arora. Map-

ping between natural movie fMRI responses and word-sequence representations. arXiv preprint arXiv:1610.03914, 2016.

6. Hejia Zhang, Po-Hsuan Chen, Janice Chen, Xia Zhu, Javier S Turek, Theodore L Willke, Uri Hasson, and Peter J Ramadge. A searchlight factor model approach for locating shared information in multi-subject fMRI analysis. arXiv preprint arXiv:1609.09432, 2016.

7. Michael J Anderson, Mihai Capot a, Javier S Turek, Xia Zhu, Theodore L Willke, Yida Wang, Po-Hsuan Chen, Jeremy R Manning, Peter J Ramadge, and Kenneth A Norman. Enabling factor analysis on thousand-subject neuroimaging datasets. arXiv preprint arXiv:1608.04647, 2016.

Bibliography

[1] Alexandre Abraham, Elvis Dohmatob, Bertrand Thirion, Dimitris Samaras, and Gael Varoquaux. Extracting brain regions from rest fMRI with total-variation constrained dictionary learning. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 607–615. Springer, 2013.

[2] Jong-Hoon Ahn and Jong-Hoon Oh. A constrained EM algorithm for principal component analysis. *Neural Computation*, 15(1):57–65, 2003.

[3] Shotaro Akaho. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071*, 2006.

[4] Hirotugu Akaike. Canonical correlation analysis of time series and the use of an information criterion. *Mathematics in Science and Engineering*, 126:27–96, 1976.

[5] Rami Al-Rfou, Guillaume Alain, Amjad Almahairi, Christof Angermueller, Dzmitry Bahdanau, Nicolas Ballas, Frédéric Bastien, Justin Bayer, Anatoly Belikov, Alexander Belopolsky, et al. Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*, 2016.

[6] Daniel L Ames, Christopher J Honey, Michael A Chow, Alexander Todorov, and Uri Hasson. Contextual alignment of cognitive and neural dynamics. *Journal of cognitive neuroscience*, 2014.

[7] Michael J Anderson, Mihai Capotă, Javier S Turek, Xia Zhu, Theodore L Willke, Yida Wang, Po-Hsuan Chen, Jeremy R Manning, Peter J Ramadge, and Kenneth A Norman. Enabling factor analysis on thousand-subject neuroimaging datasets. *arXiv preprint arXiv:1608.04647*, 2016.

[8] Galen Andrew, Raman Arora, Jeff A Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International Conference on Machine Learning*, pages 1247–1255, 2013.

[9] Cédric Archambeau and Francis R Bach. Sparse probabilistic projections. In *Advances in Neural Information Processing Systems*, pages 73–80, 2009.

[10] Raman Arora and Karen Livescu. Multi-view CCA-based acoustic features for phonetic recognition across speakers and domains. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7135–7139, 2013.

[11] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, 2017.

[12] Francis R Bach and Michael I Jordan. Kernel independent component analysis. *Journal of machine learning research*, 3(Jul):1–48, 2002.

[13] Francis R Bach and Michael I Jordan. A probabilistic interpretation of canonical correlation analysis. 2005.

[14] Andrew R Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

[15] Yoshua Bengio. Learning deep architectures for AI. *Foundations and trends in Machine Learning*, 2(1):1–127, 2009.

[16] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

[17] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[18] Matthew Brett, Ingrid S Johnsrude, and Adrian M Owen. The problem of functional localization in the human brain. *Nature reviews neuroscience*, 3(3):243–249, 2002.

[19] Danilo Bzdok, Michael Eickenberg, Olivier Grisel, Bertrand Thirion, and Gaël Varoquaux. Semi-supervised factored logistic regression for high-dimensional neuroimaging data. In *Advances in Neural Information Processing Systems*, pages 3348–3356, 2015.

[20] Vince D Calhoun, Tulay Adali, Godfrey D Pearlson, and JJ Pekar. A method for making group inferences from functional MRI data using independent component analysis. *Human brain mapping*, 14(3):140–151, 2001.

[21] Vince D Calhoun, Jingyu Liu, and Tülay Adalı. A review of group ICA for fMRI data and ICA for joint inference of imaging, genetic, and ERP data. *Neuroimage*, 45(1):S163–S172, 2009.

[22] Vince D Calhoun, Rogers F Silva, Tülay Adalı, and Srinivas Rachakonda. Comparison of PCA approaches for very large group ICA. *NeuroImage*, 118:662–666, 2015.

[23] Sarath Chandar, Mitesh M Khapra, Hugo Larochelle, and Balaraman Ravindran. Correlational neural networks. *Neural computation*, 2015.

[24] Kamalika Chaudhuri, Sham M Kakade, Karen Livescu, and Karthik Sridharan. Multi-view clustering via canonical correlation analysis. In *International Conference on Machine Learning*, pages 129–136. ACM, 2009.

[25] Janice Chen, Yuan Chang Leong, Christopher J Honey, Chung H Yong, Kenneth A Norman, and Uri Hasson. Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, 20(1):115–125, 2017.

[26] Jianmin Chen, Rajat Monga, Samy Bengio, and Rafal Jozefowicz. Revisiting distributed synchronous SGD. *arXiv preprint arXiv:1604.00981*, 2016.

[27] Po-Hsuan Chen, Janice Chen, Yaara Yeshurun, Uri Hasson, James Haxby, and Peter J Ramadge. A reduced-dimension fMRI shared response model. In *Advances in Neural Information Processing Systems*, pages 460–468, 2015.

[28] Po-Hsuan Chen, J Swaroop Guntupalli, James V Haxby, and Peter J Ramadge. Joint SVD-Hyperalignment for multi-subject fMRI data alignment. In *IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*, pages 1–6, 2014.

[29] Po-Hsuan Chen, Xia Zhu, Hejia Zhang, Javier S Turek, Janice Chen, Theodore L Willke, Uri Hasson, and Peter J Ramadge. A convolutional autoencoder for multi-subject fMRI data aggregation. *arXiv preprint arXiv:1608.04846*, 2016.

[30] Xi Chen, Han Liu, and Jaime G Carbonell. Structured sparse canonical correlation analysis. In *AISTATS*, volume 12, pages 199–207, 2012.

[31] Jonathan D Cohen, Nathaniel Daw, Barbara Engelhardt, Uri Hasson, Kai Li, Yael Niv, Kenneth A Norman, Jonathan Pillow, Peter J Ramadge, Nicholas B Turk-Browne, et al. Computational approaches to fMRI analysis. *Nature Neuroscience*, 20(3):304–313, 2017.

[32] Bryan Conroy, Ben Singer, James Haxby, and Peter J Ramadge. fMRI-based inter-subject cortical alignment using functional connectivity. In *Advances in Neural Information Processing Systems*, pages 378–386, 2009.

[33] Bryan R Conroy, Benjamin D Singer, J Swaroop Guntupalli, Peter J Ramadge, and James V Haxby. Inter-subject alignment of human cortical anatomy using functional connectivity. *NeuroImage*, 81:400–411, 2013.

[34] MeganT. deBettencourt, Jonathan D. Cohen, Ray F. Lee, Kenneth A. Norman, and Nicholas B. Turk-Browne. Closed-loop training of attention with real-time brain imaging. *Nature neuroscience*, 18(3):470–475, 2015.

[35] Paramveer Dhillon, Dean P Foster, and Lyle H Ungar. Multi-view learning of word embeddings via CCA. In *Advances in Neural Information Processing Systems*, pages 199–207, 2011.

[36] Elvis Dohmatob, Arthur Mensch, Gael Varoquaux, and Bertrand Thirion. Learning brain regions via large-scale online structured sparse dictionary learning. In *Advances in Neural Information Processing Systems*, pages 4610–4618, 2016.

[37] Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[38] Michael Eickenberg, Elvis Dohmatob, Bertrand Thirion, and Gaël Varoquaux. Grouping total variation and sparsity: Statistical learning with segmenting penalties. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 685–693. Springer International Publishing, 2015.

[39] Ali Mamdouh Elkahky, Yang Song, and Xiaodong He. A multi-view deep learning approach for cross domain user modeling in recommendation systems. In *Proceedings of the 24th International Conference on World Wide Web*, pages 278–288. ACM, 2015.

[40] Erik Barry Erhardt, Srinivas Rachakonda, Edward J Bedrick, Elena A Allen, Tülay Adali, and Vince D Calhoun. Comparison of multi-subject ICA methods for analysis of fMRI data. *Human brain mapping*, 32(12):2075–2095, 2011.

[41] Joset A Etzel, Jeffrey M Zacks, and Todd S Braver. Searchlight analysis: promise, pitfalls, and potential. *Neuroimage*, 78:261–269, 2013.

[42] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482, 2015.

[43] Orhan Firat, Emre Aksan, Ilke Oztekin, and Fatos T Yarman Vural. Learning deep temporal representations for brain decoding. *arXiv preprint arXiv:1412.7522*, 2014.

[44] Bruce Fischl, Martin I Sereno, Roger BH Tootell, Anders M Dale, et al. High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, 8(4):272–284, 1999.

[45] Michael D Fox and Marcus E Raichle. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nature Reviews Neuroscience*, 8(9):700–711, 2007.

[46] Yusuke Fujiwara, Yoichi Miyawaki, and Yukiyasu Kamitani. Estimating image bases for visual image reconstruction from human brain activity. In *Advances in neural information processing systems*, pages 576–584, 2009.

[47] John C Gower and Garmt B Dijksterhuis. *Procrustes problems*. Number 30. Oxford University Press on Demand, 2004.

[48] Timothy D Griffiths and Jason D Warren. The planum temporale as a computational hub. *Trends in neurosciences*, 25(7):348–353, 2002.

[49] Kalanit Grill-Spector and Kevin S Weiner. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548, 2014.

[50] J Swaroop Guntupalli, Michael Hanke, Yaroslav O Halchenko, Andrew C Connolly, Peter J Ramadge, and James V Haxby. A model of representational spaces in human cortex. *Cerebral Cortex*, 2016.

[51] J Swaroop Guntupalli and James V Haxby. A computational model of shared fine-scale structure in the human connectome. *bioRxiv*, page 108738, 2017.

[52] Michael Hanke, Florian J Baumgartner, Pierre Ibe, Falko R Kaule, Stefan Pollmann, Oliver Speck, Wolf Zinke, and Jörg Stadler. A high-resolution 7-tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific data*, 1, 2014.

[53] Michael Hanke, Yaroslav O Halchenko, Per B Sederberg, Stephen José Hanson, James V Haxby, and Stefan Pollmann. Pymvpa: a python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, 7(1):37–53, 2009.

[54] Michael Hanke, Yaroslav O Halchenko, Per B Sederberg, Emanuele Olivetti, Ingo Fründ, Jochem W Rieger, Christoph S Herrmann, James V Haxby, Stephen José Hanson, and Stefan Pollmann. Pymvpa: a unifying approach to the analysis of neuroscientific data.

[55] David R Hardoon and John Shawe-Taylor. Sparse canonical correlation analysis. *Machine Learning*, 83(3):331–353, 2011.

[56] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[57] Uri Hasson, Janice Chen, and Christopher J Honey. Hierarchical process memory: memory as an integral component of information processing. *Trends in cognitive sciences*, 19(6):304–313, 2015.

[58] Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664):1634–1640, 2004.

[59] James V Haxby, M Ida Gobbini, Maura L Furey, Alumit Ishai, Jennifer L Schouten, and Pietro Pietrini. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293(5539):2425–2430, 2001.

[60] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011.

[61] Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

[62] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[63] Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.

[64] Scott A Huettel, Allen W Song, and Gregory McCarthy. *Functional magnetic resonance imaging*, volume 1. Sinauer Associates Sunderland, 2004.

[65] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

[66] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012.

[67] Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component analysis*, volume 46. John Wiley & Sons, 2004.

[68] Aapo Hyvärinen and Erkki Oja. Independent component analysis: algorithms and applications. *Neural networks*, 13(4), 2000.

[69] Forrest N Iandola, Matthew W Moskewicz, Khalid Ashraf, and Kurt Keutzer. Firecaffe: near-linear acceleration of deep neural network training on compute clusters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2592–2600, 2016.

[70] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, 2008.

[71] Jon R Kettenring. Canonical analysis of several sets of variables. *Biometrika*, pages 433–451, 1971.

[72] Einat Kidron, Yoav Y Schechner, and Michael Elad. Pixels that sound. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 88–95, 2005.

[73] George Kimeldorf and Grace Wahba. Some results on tchebycheffian spline functions. *Journal of mathematical analysis and applications*, 33(1):82–95, 1971.

[74] Ryan Kiros, Ruslan Salakhutdinov, and Richard S Zemel. Multimodal neural language models. In *International Conference on Machine Learning*, volume 14, pages 595–603, 2014.

[75] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Skip-thought vectors. In *Advances in neural information processing systems*, pages 3294–3302, 2015.

[76] Sotetsu Koyamada, Yumi Shikauchi, Ken Nakae, Masanori Koyama, and Shin Ishii. Deep learning of fMRI big data: a novel approach to subject-transfer decoding. *arXiv preprint arXiv:1502.00093*, 2015.

[77] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National academy of Sciences of the United States of America*, 103(10):3863–3868, 2006.

[78] Li-Dan Kuang, Qiu-Hua Lin, Xiao-Feng Gong, Fengyu Cong, Jing Sui, and Vince D Calhoun. Multi-subject fMRI analysis via combined independent component analysis and shift-invariant canonical polyadic decomposition. *Journal of neuroscience methods*, 256:127–140, 2015.

[79] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, volume 14, pages 1188–1196, 2014.

[80] Jong-Hwan Lee, Te-Won Lee, Ferenc A Jolesz, and Seung-Schik Yoo. Independent vector analysis (IVA): multivariate approach for fMRI group study. *Neuroimage*, 40(1):86–109, 2008.

[81] Yi-Ou Li, Tülay Adali, Wei Wang, and Vince D Calhoun. Joint blind source separation by multiset canonical correlation analysis. *IEEE Transactions on Signal Processing*, 57(10):3918–3929, 2009.

[82] Nikos K Logothetis. The neural basis of the blood–oxygen–level–dependent functional magnetic resonance imaging signal. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 357(1424):1003–1037, 2002.

[83] Alexander Lorbert and Peter J Ramadge. Kernel hyperalignment. In *Advances in Neural Information Processing Systems*, pages 1790–1798, 2012.

[84] Jeremy R Manning, Rajesh Ranganath, Waitsang Keung, Nicholas B Turk-Browne, Jonathan D Cohen, Kenneth A Norman, and David M Blei. Hierarchical topographic factor analysis. In *IEEE International Workshop on Pattern Recognition in Neuroimaging*, pages 1–4, 2014.

[85] Jeremy R Manning, Rajesh Ranganath, Kenneth A Norman, and David M Blei. Topographic factor analysis: a bayesian model for inferring brain networks from neural data. *PloS one*, 9(5):e94914, 2014.

[86] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015.

[87] Daniel S Margulies, Justin L Vincent, Clare Kelly, Gabriele Lohmann, Lucina Q Uddin, Bharat B Biswal, Arno Villringer, F Xavier Castellanos, Michael P Milham, and Michael Petrides. Precuneus shares intrinsic functional architecture in humans and monkeys. *Proceedings of the National Academy of Sciences*, 106(47):20069–20074, 2009.

[88] John Mazziotta, Arthur Toga, Alan Evans, Peter Fox, Jack Lancaster, Karl Zilles, Roger Woods, Tomas Paus, Gregory Simpson, Bruce Pike, et al. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 356(1412):1293–1322, 2001.

[89] Thomas Melzer, Michael Reiter, and Horst Bischof. Nonlinear feature extraction using generalized canonical correlation analysis. In *International Conference on Artificial Neural Networks*, pages 353–360. Springer, 2001.

[90] Arthur Mensch, Gaël Varoquaux, and Bertrand Thirion. Compressed online dictionary learning for fast resting-state fMRI decomposition. In *Biomedical Imaging (ISBI), 2016 IEEE 13th International Symposium on*, pages 1282–1285. IEEE, 2016.

[91] Andrew M Michael, Mathew Anderson, Robyn L Miller, Tülay Adalı, and Vince D Calhoun. Preserving subject variability in group fMRI analysis: performance evaluation of GICA vs. IVA. *Frontiers in systems neuroscience*, 8, 2014.

[92] Tom M Mitchell, Svetlana V Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L Malave, Robert A Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008.

[93] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410, 2011.

[94] Andrew Ng. Sparse autoencoder. *CS294A Lecture notes*, 2011.

[95] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng. Multimodal deep learning. In *International Conference on Machine Learning*, pages 689–696, 2011.

[96] Shinji Nishimoto, An T Vu, Thomas Naselaris, Yuval Benjamini, Bin Yu, and Jack L Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 21(19):1641–1646, 2011.

[97] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9):424–430, 2006.

[98] Elena Parkhomenko, David Tritchler, Joseph Beyene, et al. Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–34, 2009.

[99] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Empirical Methods on Natural Language Processing*, volume 14, pages 1532–1543, 2014.

[100] Francisco Pereira, Greg Detre, and Matthew Botvinick. Generating text from functional brain images. *Frontiers in human neuroscience*, 5:72, 2011.

[101] Francisco Pereira, Bin Lou, Brianna Pritchett, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. Decoding of generic mental representations from functional mri data using word embeddings. *bioRxiv*, page 057216, 2016.

[102] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, 45(1):S199–S209, 2009.

[103] Sergey M Plis, Devon R Hjelm, Ruslan Salakhutdinov, and Vince D Calhoun. Deep learning for neuroimaging: a validation study. *Frontiers in Neuroscience*, 8:229, 2014.

[104] Russell A Poldrack. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron*, 72(5):692–697, 2011.

[105] Marcus E Raichle. The brain's default mode network. *Annual review of neuroscience*, 38:433–447, 2015.

[106] Mor Regev, Christopher J Honey, Erez Simony, and Uri Hasson. Selective and invariant neural responses to spoken and written narratives. *Journal of Neuroscience*, 33(40):15978–15988, 2013.

[107] Jan Rupnik, Primoz Skraba, John Shawe-Taylor, and Sabrina Guettes. A comparison of relaxations of multiset cannonical correlation analysis and applications. *arXiv preprint arXiv:1302.0974*, 2013.

[108] Mert R Sabuncu, Benjamin D Singer, Bryan Conroy, Ronald E Bryan, Peter J Ramadge, and James V Haxby. Function-based intersubject alignment of human cortical anatomy. *Cerebral Cortex*, 20(1):130–140, 2010.

[109] Bernhard Schölkopf, Ralf Herbrich, and Alex J Smola. A generalized representer theorem. In *International Conference on Computational Learning Theory*, pages 416–426. Springer, 2001.

[110] Jonathan Richard Shewchuk et al. An introduction to the conjugate gradient method without the agonizing pain, 1994.

[111] Erez Simony, Christopher J Honey, Janice Chen, Olga Lositsky, Yaara Yeshurun, Ami Wiesel, and Uri Hasson. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature communications*, 7, 2016.

[112] Stephen M Smith. The future of fMRI connectivity. *Neuroimage*, 62(2):1257–1266, 2012.

[113] Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, 2014.

[114] Jean Talairach and Pierre Tournoux. Co-planar stereotaxic atlas of the human brain 3-dimensional proportional system: an approach to cerebral imaging. 1988.

[115] Bertrand Thirion, Edouard Duchesnay, Edward Hubbard, Jessica Dubois, Jean-Baptiste Poline, Denis Lebihan, and Stanislas Dehaene. Inverse retinotopy: inferring the visual content of images from brain activation patterns. *Neuroimage*, 33(4):1104–1116, 2006.

[116] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2), 2012.

[117] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[118] Roger BH Tootell, John B Reppas, Anders M Dale, Rodney B Look, et al. Visual motion aftereffect in human cortical area mt revealed by functional magnetic resonance imaging. *Nature*, 375(6527):139, 1995.

[119] James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *Journal of Machine Learning Research*, 17(137):1–5, 2016.

[120] Javier S. Turek, Theodore L. Willke, Po-Hsuan Chen, and Peter J. Ramadge. A semi-supervised method for multi-subject fMRI functional alignment. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017.

[121] Gaël Varoquaux, Michael Eickenberg, Elvis Dohmatob, and Bertand Thirion. Faasta: A fast solver for total-variation regularization of ill-conditioned problems with application to brain imaging. *arXiv preprint arXiv:1512.06999*, 2015.

[122] Gaël Varoquaux, Alexandre Gramfort, Fabian Pedregosa, Vincent Michel, and Bertrand Thirion. Multi-subject dictionary learning to segment an atlas of brain spontaneous activity. In *Biennial International Conference on Information Processing in Medical Imaging*, pages 562–573. Springer, 2011.

[123] Gaël Varoquaux, Sepideh Sadaghiani, Philippe Pinel, Andreas Kleinschmidt, Jean-Baptiste Poline, and Bertrand Thirion. A group model for stable multi-subject ICA on fMRI datasets. *Neuroimage*, 51(1):288–299, 2010.

[124] Hrishikesh D Vinod. Canonical ridge and econometrics of joint production. *Journal of econometrics*, 4(2):147–166, 1976.

[125] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164, 2015.

[126] Kiran Vodrahalli, Po-Hsuan Chen, Yingyu Liang, Janice Chen, Esther Yong, Christopher Honey, Peter Ramadge, Ken Norman, and Sanjeev Arora. Mapping between natural movie fMRI responses and word-sequence representations. *arXiv preprint arXiv:1610.03914*, 2016.

[127] Weiran Wang, Raman Arora, Karen Livescu, and Jeff A Bilmes. On deep multi-view representation learning. In *International Conference on Machine Learning*, pages 1083–1092, 2015.

[128] Weiran Wang, Honglak Lee, and Karen Livescu. Deep variational canonical correlation analysis. *arXiv preprint arXiv:1610.03454*, 2016.

[129] John DG Watson, Ralph Myers, Richard SJ Frackowiak, Joseph V Hajnal, Roger P Woods, John C Mazziotta, Stewart Shipp, and Semir Zeki. Area v5 of the human brain: evidence from a combined study using positron emission tomography and magnetic resonance imaging. *Cerebral cortex*, 3(2):79–94, 1993.

[130] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014.

[131] Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom M Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *EMNLP*, pages 233–243, 2014.

[132] John Wieting, Mohit Bansal, Kevin Gimpel, and Karen Livescu. Towards universal paraphrastic sentence embeddings. *International Conference on Learning Representations*, 2016.

[133] Daniela M Witten, Robert J Tibshirani, et al. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical applications in genetics and molecular biology*, 8(1):1–27, 2009.

[134] Hao Xu, Alexander Lorbert, Peter J Ramadge, J Swaroop Guntupalli, and James V Haxby. Regularized hyperalignment of multi-set fMRI data. In *IEEE Statistical Signal Processing Workshop (SSP)*, pages 229–232, 2012.

[135] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, volume 14, 2015.

[136] Fei Yan and Krystian Mikolajczyk. Deep correlation for matching images and text. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3441–3450, 2015.

[137] Yaara Yeshurun, Stephen Swanson, Janice Chen, Erez Simony, Christopher Honey, Christina Lazaridi, and Uri Hasson. How does the brain represent different ways of understanding the same story? *Society for Neuroscience Abstracts*, 2014.

[138] Yaara Yeshurun, Stephen Swanson, Erez Simony, Janice Chen, Christina Lazaridi, Christopher J Honey, and Uri Hasson. Same story, different story: The neural representation of interpretive frameworks. *Psychological Science*.

[139] Hejia Zhang, Po-Hsuan Chen, Janice Chen, Xia Zhu, Javier S Turek, Theodore L Willke, Uri Hasson, and Peter J Ramadge. A searchlight factor model approach for locating shared information in multi-subject fMRI analysis. *arXiv preprint arXiv:1609.09432*, 2016.