UNIVERSITY OF CALIFORNIA SAN DIEGO

Decomposition of Memory Using Single-Trial EEG Classifier

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Electrical Engineering (Machine Learning and Data Science)

by

Kuei-Da Liao

Committee in charge:

Professor Virginia R. de Sa, Chair
Professor Vikash Gilja, Co-Chair
Professor Marta Kutas
Professor Truong Quang Nguyen
Professor Mohan M. Trivedi

2021

The Dissertation of Kuei-Da Liao is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

To my parents and my sister, for their unending support throughout the journey.

And to Chia-Ling, for her unlimited love and unconditional encouragement.

TABLE OF CONTENTS

Dissertation Approval Page iii

Dedication iv

Table of Contents v

List of Figures viii

List of Tables xii

Acknowledgements xv

Vita xvii

Abstract of the Dissertation xviii

Chapter 1 Introduction 1
1.1 Electroencephalography 2
1.2 Methodologies to Measure Neural Responses 2

Chapter 2 Material - EEG Dataset 5
2.1 Experiment 1 5
2.1.1 Participants 5
2.1.2 Experimental Paradigm and EEG Acquisition 5
2.2 Experiment 2 8
2.2.1 Participants 8
2.2.2 Experimental Paradigm and EEG Acquisition 8
2.3 Experiment 3 10
2.3.1 Participants 10
2.3.2 Experimental Paradigm and EEG Acquisition 10
2.4 Preprocessing 11
2.5 Chapter Acknowledgements 12

Chapter 3 Predicts Memory Retrieval within Subjects 13
3.1 Introduction 13
3.2 Methods 15
3.2.1 Classification Problems 15
3.2.2 Classification 16
3.2.3 Statistical Methods 17
3.3 Results 18
3.3.1 Classifier Performance 18
3.3.2 Analysis of the Classifier Scores 22
3.3.3 Classifier Activation Patterns 24

3.3.4 Classifier Scores Evolution over Time . . . 26
3.4 Discussion . . . 28
3.5 Chapter Acknowledgements . . . 34

Chapter 4 Predicts Memory Retrieval across Subjects . . . 35
4.1 Introduction . . . 35
4.2 Methods . . . 36
4.2.1 Classification Problem . . . 36
4.2.2 Subject-Independent Classification . . . 37
4.2.3 Statistical Methods . . . 38
4.3 Results . . . 38
4.3.1 Classifier Performance . . . 38
4.3.2 Analysis of Classifier Scores . . . 40
4.3.3 Activation Patterns . . . 40
4.4 Discussion . . . 40
4.5 Chapter Acknowledgements . . . 46

Chapter 5 Control for Confidence Reveals Familiarity . . . 47
5.1 Introduction . . . 47
5.2 Methods . . . 48
5.2.1 Training Classifiers . . . 49
5.2.2 Differentiate Conditions not Trained . . . 50
5.2.3 Condition-Controlled Classifier Training . . . 51
5.2.4 Visualization of Consistent EEG Features . . . 51
5.3 Results . . . 51
5.3.1 Classifier Performance . . . 51
5.3.2 Projections from the Classifiers . . . 52
5.3.3 Activation Patterns . . . 53
5.4 Discussion . . . 54
5.4.1 Difference between SN and MN . . . 55
5.4.2 Confidence Component Revealed in F vs. CR . . . 55
5.4.3 Confidence Matched F vs. CR . . . 56
5.5 Chapter Acknowledgements . . . 56

Chapter 6 Remember-Know Responses: Difference in Confidence, Source Memory, and Item Memory . . . 58
6.1 Introduction . . . 58
6.1.1 ERPs in EEG Data Analysis . . . 58
6.1.2 EEG, Single-Trial Classifier, LOSO . . . 60
6.1.3 Our Goal . . . 61
6.2 Selected Behaviors for Training the Classifier . . . 61
6.2.1 Materials . . . 62
6.2.2 Classification Problem . . . 62
6.2.3 Methods . . . 63

6.2.4 Results 68
6.2.5 Discussion 80
6.3 Interpretation of RS vs. F using Memory and Confidence Components 86
6.3.1 Methods 86
6.3.2 Results 87
6.3.3 Discussion 94
6.4 General Discussion 101
6.4.1 Subject-Dependent vs. Subject-Independent Training 101
6.4.2 The Components for Confidence in Correct Old and New Responses Are the Same 102
6.4.3 Linear Regression Models for Decomposing Cognitive Components . . . 102
6.5 Revisiting Source Memory in Experiment 2 and 3 103
6.5.1 Methods 104
6.5.2 Results 104
6.5.3 Discussion 106
6.6 Chapter Acknowledgements 107

Chapter 7 Memory Model Consists of Source Memory, Item memory, and Confidence 109
7.1 Remember-Know (RK) Paradigm 109
7.2 Single-Process and Dual-Process Model 110
7.3 Problems of Single- and Dual-Process Models 111
7.4 Our Proposed Model 112
7.5 Chapter Acknowledgements 115

Chapter 8 Summary 116

Bibliography 118

LIST OF FIGURES

Figure 2.1. Experimental paradigms for (a) location source information and (b) color source experiments. 7

Figure 2.2. The GSN electrode layout used for EEG recording and the six channels groups on which classification analysis was conducted. 8

Figure 3.1. The ROC curves for the three different classification problems (A: SC vs. CR; B: SI vs. CR; and C: SC vs. SI) are given separately for the four individual datasets. 21

Figure 3.2. The average of the estimated means and the approximate confidence intervals of the classifier scores across the four datasets (Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col) for the five behavioral conditions for the three different classification problems. 23

Figure 3.3. The average of the estimated means and the approximate confidence intervals of the classifier scores across the four datasets (Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col) when considering the breakdown by subjective ratings for the three different classification problems. 25

Figure 3.4. The average activation patterns averaged across all available subjects from (A) the SC-CR classifiers, (B) the SI-CR classifiers, and (C) the SC-SI classifiers. 26

Figure 3.5. The average activation patterns masked by the most significant cluster for the three different classification problems (A: SC vs. CR; B: SI vs. CR; C: SC vs. SI). 27

Figure 3.6. (A) The scores of all conditions across time by SC-CR classifiers. (B) The scores of all conditions across time by SI-CR classifiers. (C) The scores of all conditions across time by SC-SI classifiers. 29

Figure 4.1. The ROC curves for the four individual datasets are given in the three different classification problems (a) SC-CR, (b) SI-CR, and (c) SC-SI. . . . 39

Figure 4.2. The scores of projected trials in different behaviors using projection functions from (a) SC-CR classifiers, (b) SI-CR classifiers, and (c) SC-SI classifiers are given separately for four individual datasets. 41

Figure 4.3. The patterns are the average of normalized mean difference between two classes for each subject. 42

Figure 4.4. The significant clusters of features ($p < .05$) in the patterns of the average of normalized mean difference between two classes for each subject. 43

Figure 5.1. Green and red box outlines indicate the positive and negative behaviors for training the SN vs. MN classifier. 49

Figure 5.2. Green and red box outlines indicate the positive and negative behaviors for training the F vs. CR classifier. 49

Figure 5.3. The average projection and the 95% CI of behaviors from the trained classifiers in Exp 1 and 2. The positive and negative trained classes are plotted in green and red, respectively. The classes shown in black were not trained. 53

Figure 5.4. Mean difference between training classes of each classification problem in Exp 1 and 2. The areas surrounded by the dotted lines are the top positive/negative clusters with p-value $< .05$. 54

Figure 6.1. CR-SN and CR-MN were selected for training the CR-SN vs. CR-MN confidence classifier. Green box and red box are for the positive and negative class, respectively. 64

Figure 6.2. SC-RS and SI-RS were selected for training the SC-RS vs. SI-RS source memory classifier. Green box and red box are for the positive and negative class, respectively. 64

Figure 6.3. SI-RS, SI-RO, SI-F and CR-SN, CR-MN were selected for training the SI vs. CR item memory classifier. Green box and red box are for the positive and negative class, respectively. 64

Figure 6.4. SC-RS, SC-F, and SI-F were selected for training the RS vs. F R-K classifier. Green box and red box are for the positive and negative class, respectively. 65

Figure 6.5. The average projection and the 95% confidence interval of behaviors from (1) CR-SN vs. CR-MN confidence classifier, (2) SC-RS vs. SI-RS source memory classifier, and (3) SC-RS vs. SI-RS source memory classifier with confidence control in (a) Exp 1 and (b) Exp 2. 73

Figure 6.6. The average projection and the 95% confidence interval of behaviors from (4) SI vs. CR item memory classifier, (5) SI vs. CR item memory classifier with confidence control, and (6) RS vs. F R-K classifier in (a) Exp 1 and (b) Exp 2. 74

Figure 6.7. Mean difference between training classes of each classification problem (1: CR-SN vs. CR-MN, 2: SC-RS vs. SI-RS, 3: SC-RS vs. SI-RS with confidence control) in (a) Exp 1 and (b) Exp 2. The areas surrounded by the dotted lines are the top positive/negative clusters with p-value $< .05$. . . 75

Figure 6.8. Mean difference between training classes of each classification problem (4: SI vs. CR, 5: SI vs. CR with confidence control, 6: RS vs. F) in (a) Experiment 1 and (b) Experiment 2. The areas surrounded by the dotted lines are the top positive/negative clusters with p-value $< .05$. 76

Figure 6.9. (a) The average projection of behaviors from CR-SN vs. CR-MN classifier in Exp 1, (b) Exp 2, (c) with item memory control in Exp 1, and (d) Exp 2. (e) The average projection of each behavior from each subject from classifiers with and without item memory control in Exp 1 and (f) Exp 2. . 80

Figure 6.10. (a) The mean differences of CR-SN vs. CR-MN in Exp 1, (b) Exp 2, (c) with item memory control in Exp 1, and (d) Exp 2. (e) and (f) are the mean differences between CR-SN vs. CR-MN without and with item memory control in Exp 1 and Exp 2, respectively. 81

Figure 6.11. (a) The AUROCs of SC-RS vs. SC&SI-F projection on RS vs. F LOSO classifier (along x-axis), predictions of CoSmIm linear regression model using classifier (along left y-axis), and projection on RS vs. F LOTO classifier (along right y-axis) of each subject in Exp 1 and (b) Exp 2. 89

Figure 6.12. (a) The average projection and confidence interval of different behaviors from CoSmIm model for RS and F in Exp 1 and (b) Exp 2. (c) The average projection of different behaviors from RS vs. F R-K classifier in Exp 1 and (d) Exp 2. 90

Figure 6.13. (a) The combined pattern of CR-SN vs. CR-MN, SC-RS vs. SI-RS with CC, and SI vs. CR with CC in Exp 1 and (b) in Exp 2. (c) Mean difference between training classes of RS vs. F in Exp 1 and (d) Exp 2. 91

Figure 6.14. (a) The correlation between the effect of the Sm (c) Im (e) Co component to CoSmIm model and source accuracy in Exp 1. (b) The correlation between the effect of the Sm (d) Im (f) Co component to CoSmIm model and source accuracy in Exp 2. 93

Figure 6.15. (a) The correlation between the effect of the Sm (c) Im (e) Co component to CoSmIm model and FA-RS responses in Exp 1. (b) The correlation between the effect of the Sm (d) Im (f) Co component to CoSmIm model and FA-RS responses in Exp 2. 95

Figure 6.16. (a) The correlation between the effect of the Sm (c) Im (e) Co component to CoSmIm model and RS judgments with SI responses in Exp 1. (b) The correlation between the effect of the Sm (d) Im (f) Co component to CoSmIm model and RS judgments with SI responses in Exp 2. 96

Figure 6.17. (a) The correlation between the effect of the Sm (c) Im (e) Co component to CoSmIm model and RO judgments with SI responses in Exp 1. (b) The correlation between the effect of the Sm (d) Im (f) Co component to CoSmIm model and RO judgments with SI responses in Exp 2. 97

Figure 6.18. (a) The correlation between the effect of the Sm (c) Im (e) Co component to CoSmIm model and source accuracy with RO responses in Exp 1. (b) The correlation between the effect of the Sm (d) Im (f) Co component to CoSmIm model and source accuracy with RO responses in Exp 2. 98

Figure 7.1. (a) Single-process and (b) dual-process models for Remember-Know paradigm. 110

Figure 7.2. Behaviors selected for training the classifiers in Chapter 6 presented in the view of (a) Single-process and (b) dual-process models for Remember-Know paradigm. 111

Figure 7.3. Illustration of our proposed model to explain memory behaviors in three components: source memory, item memory, and confidence. Locations of the behaviors are based on Exp 1. 114

LIST OF TABLES

Table 3.1. Classification results for Experiment 1. 19

Table 3.2. Classification results for Experiment 2. 20

Table 3.3. Classification results for Experiment 3. 21

Table 3.4. Classification results for Experiment 1 to 3. 22

Table 3.4. Comparison results between the classifier scores for the SC-CR classifier. . 22

Table 3.5. The uncorrected pairwise comparison results for the five behavioral conditions across the four datasets [Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col]. 24

Table 3.6. The difference between the average classifier scores for the SC and FA conditions are given in the top three rows. 33

Table 3.7. The uncorrected pairwise comparison results for the five subjective rating options across the four datasets [Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col]. 34

Table 4.1. AUROCs are given separately for different experiments and classification problems using different training paradigms. The LOTO methods trained separate classifiers for each subject using only the subject's own data. 39

Table 4.2. P-values for the most significant cluster in Figures 4.3/4.4 42

Table 4.3. Number of trials in each class used by 3 classifiers in Experiment 1. Dash represents the classifier wasn't trained due to number of trials less than 25 in either class. 44

Table 4.4. Number of trials in each class used by 3 classifiers in Experiment 2. Dash represents the classifier wasn't trained due to number of trials less than 25 in either class. 44

Table 4.5. Number of trials in each class used by 3 classifiers in Experiment 3-loc. Dash represents the classifier wasn't trained due to number of trials less than 25 in either class. 44

Table 4.6. Number of trials in each class used by 3 classifiers in Experiment 3-col. Dash represents the classifier wasn't trained due to number of trials less than 25 in either class. 45

Table 4.7. Areas under ROC curves calculated based on the scores computed from projections of behaviors with different classifiers 46

Table 5.1. AUROCs and accuracies calculated based on the scores computed from projections of behaviors from different classifiers. RS and ConfMatched refer to SC-RS and confidence matched, respectively. 52

Table 6.1. Exp 1 ACCs of test subjects with at least 5 trials in each class. 69

Table 6.2. Exp 1 AUROCs of test subjects with at least 5 trials in each class. 70

Table 6.3. Exp 2 ACCs of test subjects with at least 5 trials in each class. 71

Table 6.4. Exp 2 AUROCs of test subjects with at least 5 trials in each class. 72

Table 6.5. Areas under ROC curves calculated based on the projections of behavior-pairs onto different classifiers 73

Table 6.6. Areas under ROC curves calculated based on the projections of behavior-pairs onto different classifiers 79

Table 6.7. Wilcoxon rank sum test on the AUROCs of subjects from CR-SN vs. CR-MN classifier before and after item memory control. 79

Table 6.8. Total numbers of trials of correct item responses in both experiments. 83

Table 6.9. The average AUROCs of classifying SC-RS and SC&SI-F using different classifiers and linear regression models. 88

Table 6.10. The p-values of the paired one-tailed Student's t-test between subject AUROCs of different linear regression models. 89

Table 6.11. The average coefficients and ratios of predictor variables in different linear regression models. 92

Table 6.12. The significance of the correlation between AUROC of CoSmIm model subtracted by AUROC of other models with two component and behavior ratio different from zero. 94

Table 6.13. The average AUROCs of classifying SC-RS and SC&SI-F using different classifiers and linear regression models with components from Exp 1 (shown as Linreg (1)). 105

Table 6.14. The p-values of the paired one-tailed Student's t-test between subject AUROCs of different linear regression models with components from Exp 1. 105

Table 6.15. The average coefficients and ratios of predictor variables in different linear regression models using components from Exp 1 (shown as CoSmIm (1)). . 106

Table 6.16. Total numbers of trials of correct item responses in Exp 3-loc and 3-col. . . . 107

ACKNOWLEDGEMENTS

First and foremost, praises and thanks to the Lord, the Almighty, for His showers of blessings and unconditional love throughout my research work, especially those most struggling days, to complete the work successfully.

I would like to express my deep and sincere gratitude to my advisor, Professor Virginia de Sa, for her invaluable guidance and support in immeasurable ways. It was a great privilege and honor to work and learn under her guidance, and I am extremely grateful for what she has offered me.

I would like to thank Professor Tim Curran and Marta Kutas for sharing their knowledge and insights on memory processes during retrieval. Their suggestions and comments have helped me find new angles and directions for memory EEG analysis.

I would also like to thank my committee: Professor Gilja, Professor Nguyen, and Professor Trivedi for generously giving their time and providing guidance and and advice.

I would like to thank my parents for their love and caring. The opportunity of education you provided was the best gift what I could have ask for. Also, many thanks to my sister and my girlfriend for their constant encouragement and love.

I am extremely grateful for having my friends and research colleagues from desalab, SCCN. There were countless times that I sought for help or advice, and they never ever failed me. My research would not be the same without their support.

Chapter 2, in part, is a reprint of the material as it appears in Frontiers in Human Neuroscience, 2018. Eunho Noh; Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. “Single-trial EEG analysis predicts memory retrieval and reveals source-dependent differences”, Frontiers in Human Neuroscience, 12, 258, 2018. The dissertation author was the primary co-author of this paper.

Chapter 3, in full, includes the result section and discussion section in the following publication in Frontiers in Human Neuroscience, 2018. Eunho Noh; Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. “Single-trial EEG analysis predicts memory retreival

and reveals source-dependent differences", Frontiers in Human Neuroscience, 12, 258, 2018. The dissertation author was the primary co-author of this publication.

Chapter 4, in full, is a reorganized version of the following publication in IEEE International Conference on Bioinformatics and Biomedicine, 2018. Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. "Single-trial EEG predicts memory retrieval using leave-one-subject-out classification", 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2613-2620, 2018. The dissertation author was the primary author of this publication.

Chapter 5, in full, is a reorganized version of the following publication in the Proceedings of the Annual Meeting of the Cognitive Science Society, 2021. Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. "EEG reveals familiarity by controlling confidence in memory retrieval", Proceeding of the Annual Meeting of the Cognitive Science Society, 43, 2021. The dissertation author was the primary author of this publication.

Chapter 6, in full, is the reorganized version of the manuscript in preparation. Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. "Using single-trial EEG classifiers to decompose remember-know difference", *in preparation*. The dissertation author is the primary author of this manuscript.

Chapter 7, in part, is the reorganized version of the material prepared for submission for publication. Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. "Using single-trial EEG classifiers to decompose remember-know difference", *in preparation*. The dissertation author is the primary author of this manuscript.

VITA

2008–2012 Bachelor of Science in Electrical Engineering
National Taiwan University

2012–2014 Master of Science in Biomedical Electronics and Bioinformatics
National Taiwan University

2014–2016 Substitute Military Service, Taiwan

2021 Doctor of Philosophy in Electrical Engineering (Machine Learning and Data Science)
University of California San Diego

PUBLICATIONS

"Single-trial EEG analysis predicts memory retrieval and reveals source-dependent differences," Frontiers in human neuroscience, vol. 12, pp 258, July 2018
"Single-trial EEG predicts memory retrieval using leave-one-subject-out classification," in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), pp 2613-2620, December 2018
"EEG Reveals Familiarity by Controlling Confidence in Memory Retrieval," in Proceedings of the Annual Meeting of the Cognitive Science Society, 2021

FIELDS OF STUDY

Major Field: Electrical Engineering (Machine Learning and Data Science)

Studies in Medical Devices & Systems
Studies in Machine Learning & Data Science
Professors Virginia R. de Sa

ABSTRACT OF THE DISSERTATION

Decomposition of Memory Using Single-Trial EEG Classifier

by

Kuei-Da Liao

Doctor of Philosophy in Electrical Engineering (Machine Learning and Data Science)

University of California San Diego, 2021

Professor Virginia R. de Sa, Chair
Professor Vikash Gilja, Co-Chair

We present the results using single-trial analyses and pattern classifier to analyze Electroencephalography (EEG) data recorded during recognition memory experiments.

In Chapter 2, the details of the recorded data and the experimental paradigms for both location source information and frame color source information experiments were given, and the data recorded would be explored throughout the entire study.

In Chapter 3, we used subject-dependent leave-one-trial-out (LOTO) pattern classifiers to extract features related to recognition memory retrieval from the spatio-temporal information in single-trial EEG data during attempted memory retrieval. The results showed that location

may be bound more tightly with the item than an extrinsic color association. The multivariate classification approach also showed that trial-by-trial variation in EEG corresponding to these ERP components were predictive of subjects' behavioral responses.

In Chapter 4, we performed EEG classification in memory retrieval predictions using classifiers trained on a leave-one-subject-out (LOSO) cross-validation basis. We also compared the performance to that of classification using LOTO when trained on data for an individual subject. Unlike traditional single-trial EEG analysis performed within an individual subject, we show that it is possible to perform single-trial EEG classification using classifiers trained on different subjects.

In Chapter 5, we explored the separation of decision confidence and familiarity components in EEG data from recognition memory experiments. We first developed and tested a classifier designed to classify decision confidence on new trials. We then used this classifier to control for confidence in the selection of trials of familiarity and correct rejection. This allowed us to reveal a familiarity component that is of similar magnitude for recollection and familiarity judgements. This familiarity component revealed more of a frontal extent than obtained without confidence matching.

While confidence is often considered to be indexed by memory strength, in Chapter 6, we firstly showed that the confidence classifier trained with new item decision could be disassociate with the memory strength. Moreover, projecting old decisions onto the confidence classifier revealed the same confidence feature for both the old and new item decision and that the late positive component (LPC) could be related to both confidence in old and new responses. By using linear regression models, we also showed that the difference between remember and know judgments in EEG recordings can be expressed by the difference in source memory, item memory and confidence.

In Chapter 7, we reviewed the single- and dual-process model for memory retrieval and showed how they both failed to explain the results from our trained CR-SN vs. CR-MN classifier. We then introduced our proposed model based on source memory, item memory, and confidence.

Chapter 1

Introduction

With around 100 billion neurons, the human brain is of no doubt one of the most complex system on Earth. While the neuroscientists have been dedicated to elucidating brain functions for over 100 years, we are no where close to understand well enough how the system works, e.g. processes information, orchestrates mental functions. Recent development in new neuroimaging techniques lets us visualize remarkable changes in brain. Among all different neuroimaging modalities, EEG provides the best temporal resolution with approachable budget but without intricate setup and cumbersome like functional magnetic resonance imaging (fMRI) or megnetoencephalogram (MEG). As a result, EEG has been the most widely used brain imaging technique.

One of the most interesting field in cognitive science is memory. Stretching back at least 2,000 years, Aristotle attempted to understand human memory in his treatise "On the Soul". However, the first scientific approach to study memory did not exist until the German philosopher Herman Ebbinghaus developed it in mid-1880s, where he classified memory into three categories: sensory, short-term, and long-term. With the advance of technology providing biological basis in the field of neuropsychology in the 1940s, Karl Lashley concluded that memory traces are widely distributed throughout the cortex instead of one part of the brain. In 1972, Endel Tulving first proposed two distinct kind of long-term memory, episodic and semantic. During the 1980s and 1990s, the use of computers facilitated researchers to analyze collected data during memory

experiments and propose several formal models of memory. Nowadays, the study of human memory is considered part of the disciplines of cognitive psychology and neuroscience. The improvement of neuroimaging techniques has greatly helped bridging the gap between cognitive psychology and neuroscience in memory study. EEG is one of the most popular neuroimaging techniques for memory study because of its high temporal resolution and easy accessibility.

Our research was motivated by using EEG in memory retrieval to understand how the human brain reacted to retrieve episodic memory. Below, we will provide more details of EEG during memory retrieval.

1.1 Electroencephalography

EEG is a noninvasive recording of electrical patterns on the scalp, and the majority of the collected electrical patterns in EEG is generated by groups of pyramidal neurons. When brain cells communicate with each other, they release inhibitory or excitatory neurotransmitters from the presynaptic terminal to the postsynaptic cells and generate postsynaptic potentials. With a large group of pyramidal cells aligned in parallel and excited/inhibited approximately the same time, the summation of the electric potentials could be large enough to propagate to scalp and recorded by EEG. Such electric potential could be represented as a field with positive or negative dipole. The activity from deep sources is more difficult to detect than the source near the scalp because the potential attenuates with the square of the propagated distance. The electric potential generated by the neurons in the brain becomes blurred as it spreads out through the conductor and the high resistance of the the skull. As a result, the recorded EEG on the scalp is not able to provide high spatial resolution of brain sources.

1.2 Methodologies to Measure Neural Responses

EEG is extensively used in cognitive neuroscience studies and related fields. Neural responses associated with specific sensory, motor operation, and cognitive events could be

revealed in EEG. In a standard EEG experiments, subjects are given a number of stimuli for different experimental conditions (i.e. type of stimulus presented, type of response elicited). The EEG recording in the time window corresponding to each stimulus is called a trial. Event-related potentials (ERPs) is the most common and popular method to analyze the EEG signal. ERPs are computed by averaging the EEG data across the trials of a selected stimulus or event. Because the effects of random noises could be reduced when taking average across trials, ERP analysis reveals the time- and phase-locked brain response (commoly referred to as evoked responses) to a stimulus, which would not be easily visible in a single trial of EEG in the presence of noises.

While the synchronization and desynchronization between different neurons are of interest in many EEG studies, ERP could only reflect the synchronization of the brain because ERP is the average of all frequency ranges. Also, ERP shows higher power in low frequencies but limited high frequency components in EEG because the average calculation is analogous to a low-pass filter.

Event-related desynchronization (ERD) designates a short-lasting and localized amplitude attenuation in EEG signal within a certain frequency band. On the other hand, event-related synchronization (ERS) describes a short-lasting amplitude enhancement. The oscillatory activity embedded in the brain responses could be extracted by various time-frequency analysis methods. The Fourier transform decomposes a signal depending in time domain into a spectrum in frequency domain but loses the temporal information of the oscillatory activity in the signal. The short time Fourier transform performs Fourier transform for small time windows sliding along the signal, and the temporal information of the oscillatory activity could be kept, but the limited number of points for Fourier transform would reduce the frequency resolution. Wavelet analysis gives the time-frequency decomposition using a given signal as the template and makes it capable of showing the nonstationarity in EEG signal. These time-frequency analyses allow us to examine responses, which is referred to as induced responses, which are not precisely phase-locked to an event. In fact, induced responses would usually be cancelled out by the summation due to difference in phase.

In addition to frequency analysis in temporal information, the scalp distribution of the spectral power of a given subband could also be transformed into frequency domain based on different electrode positions and analyzed. Such analysis could sometimes help hypothesize the source of the effect, although the comparably lower spatial resolution of EEG makes it less accurate than using other neuroimaging modalities like fMRI or MEG.

Chapter 2

Material - EEG Dataset

Electroencephalography recordings for the current study came from three separate visual memory task experiments Mollison and Curran (2012). All procedures were approved by the Institutional Review Board at the University of Colorado Boulder and were conducted in accordance with this approval. The written informed consent was given to all participants before the experiment. The following subsections Section 2.1, Section 2.2, and Section 2.3 mainly duplicated the details of the experimental paradigm and EEG acquisition from Mollison and Curran (2012).

2.1 Experiment 1

2.1.1 Participants

The subjects were University of Colorado undergraduate students ranging in age from 18 to 28 years old (mean age = 21.4) who volunteered for course credit (17 male, 13 female) or paid participation ($15 per hour). All subjects were right-handed native English speakers and had normal or corrected-to-normal vision.

2.1.2 Experimental Paradigm and EEG Acquisition

The experiment was divided into four blocks consisting of a study and recognition phase. The stimuli were color images of physical objects, animals, and people. There were 1297 images

in total selected from http://www.clipart.com, the stimuli set by (Brady, Konkle, Alvarez, & Oliva, 2008) and image search on the Internet. All images were resized to 240 $\times$ 240 pixels and presented on a square white background. For each subject, a total of 416 images were randomly selected as the study items, resulting in 104 items per block. The test lists comprised 100 old items from the preceding study list with 50 foil items given in random order. The first and last two stimuli in the study list were excluded from the test list to reduce primacy and recency effects.

During the study phase, the study items were presented on either the left or right side of the fixation cross. The subjects were instructed to memorize both each study item and the side of the screen on which each study item was given. The spatial location of the item was considered as the source information in this experiment. A study item was shown for 1000 ms followed by an inter-stimulus interval (ISI) of varying length (uniformly distributed within 625 $\pm$ 125 ms). A visual Gaussian noise image was shown on both the left and right sides whenever an item was not being presented to prevent after-image effects from the stimulus. The area containing the possible study image locations subtended a visual angle of 11.4° wide $\times$5.6° high.

In the recognition phase, a fixation cross appeared on the center of the screen for 750 ms. A test probe was presented for 750 ms on top of the fixation cross and followed by a 1500 ms long fixation cross. The visual angle of each test probe image was 4.3° wide $\times$4.3° high. Then the subjects were given two consecutive questions with an ISI of 625 $\pm$ 125 ms following each response. The first question had three options: left, right (given as L and R, respectively) and a new judgment (given as N). If the subjects recognize the test probe as the study item and responded with L or R as the remembered source information in the first question, they were then asked to give a modified R-K judgment in the second question. There were three options in the R-K judgment question: remember side (given as RS), remember other (given as RO), and familiar (given as F). Subjects were instructed to answer with RS if they remembered the source information of the probe, RO if they remembered information other than the source information of the probe, and F if they could not recollect any detail of the probe but it looked familiar. If a

new judgment was made in the first question, the subjects were then asked how confident they were about the test probe being a new item: sure (given as S) or maybe (given as M). Figure 2.1 (a) illustrates the study and recognition phases in the experiment. The keys for the left responses and the right responses were assigned to the left hand (z or x key) and right hand (. or / key), respectively. The key for the new response were assigned to one of the outermost keys (z or / key). For the modified R-K judgment and the confidence rating of the new item, the keys were set up from left to right according to the memory strength in either ascending or descending order. The remember (RS/RO) responses and the familiar (F) responses were always assigned to different hands. All possible key combinations were distributed to an equal number of subjects. For a given subject, the key assignment was fixed throughout the experiment.

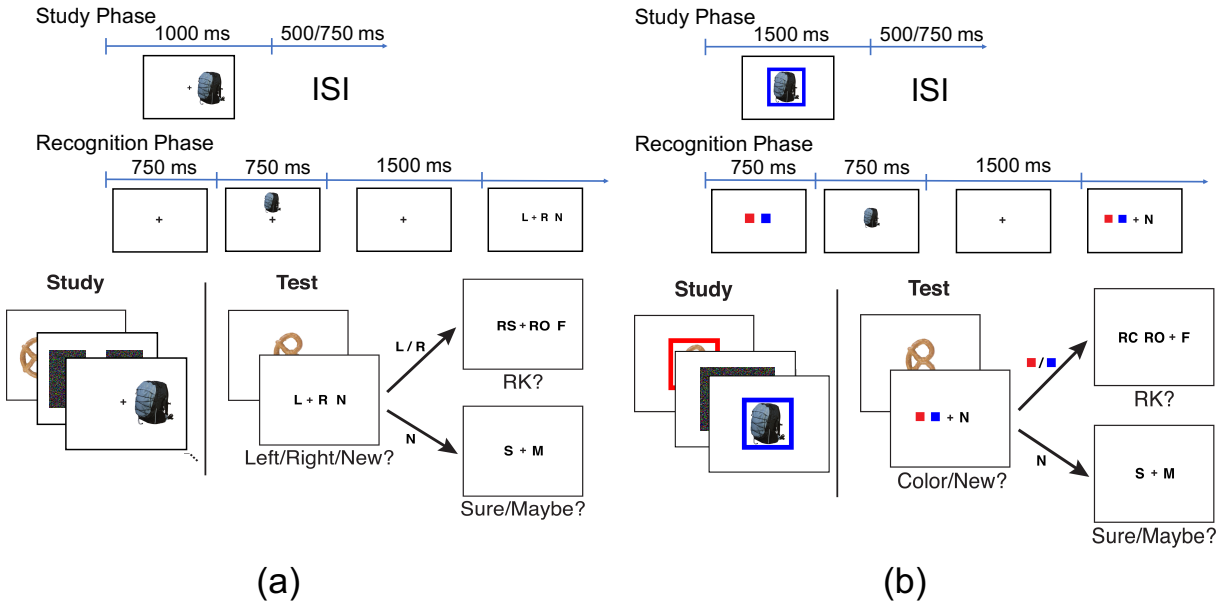


Figure 2.1. (a) Experimental paradigm for the location source information experiments (Experiment 1 and location source information sessions in Experiment 3). (b) Experimental paradigm for the color source information experiments (Experiment 2 and color source information sessions in Experiment 3).

A 128-channel Geodesic Sensor Net^{TM} [HydroCel GSN 200, v.2.1; Tucker (1993)] was used for EEG recording with the vertex a central vertex reference (Cz) and a 0.1-100 Hz bandpass hardware filter at 250 Hz sampling rate. The net was connected to an AC-coupled 128-channel,

high-input impedance amplifier (300 MΩ, Net Amps^{TM}; Electrical Geodesics Inc., Eugene, OR, United States). The impedance measurements of the electrodes were adjusted to lower than 40 kΩ. Figure 2.2 shows the locations of the electrodes.

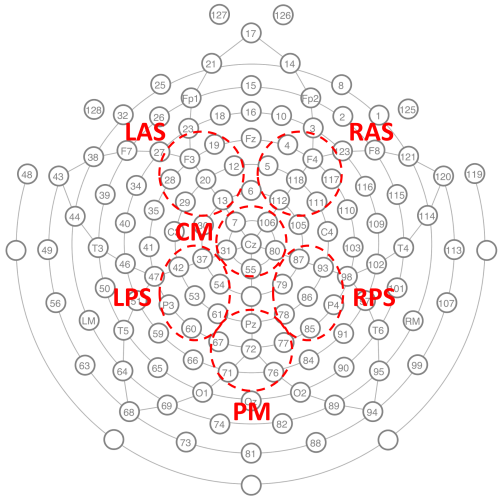


Figure 2.2. The GSN electrode layout used for EEG recording and the six channels groups on which classification analysis was conducted. Left anterior superior was given as LAS, right anterior superior as RAS, central medial as CM, left posterior superior as LPS, right posterior superior as RPS, and posterior medial as PM.

2.2 Experiment 2

2.2.1 Participants

The subjects were University of Colorado undergraduate students ranging in age from 18 to 27 years old (mean age = 21.2) who volunteered for course credit (17 male, 13 female) or paid participation ($15 per hour). All subjects were right-handed native English speakers and had normal or corrected-to-normal vision.

2.2.2 Experimental Paradigm and EEG Acquisition

The stimuli sets used in Experiments 1 and 2 were identical. In the study phase, the study item was presented and surrounded by a 48-pixel wide frame in eight possible colors (blue,

brown, green, orange, pink, purple, red, and yellow). The color of the surrounding frame was regarded as the source information in this experiment. There were four blocks consisting of a study and recognition phase. Four study lists were distributed to each study block: two used six colors and two others used the two remaining colors. Half of the subjects received the two-color study lists in odd blocks and the other half of the subjects received the two-color study lists in the even blocks. All colors were randomly and evenly selected over the study items.

During the study phase, each study item along with the frame were presented on the center of the screen. The subjects were instructed to remember each of the given study items and the corresponding frame color. A study item was shown for 1500 ms followed by an ISI with varying length (uniformly distributed within 625 ± 125 ms). A visual Gaussian noise image was shown at the location of the study item presentation whenever an item was not being presented to prevent after-image effects from the stimulus. The area containing the study image subtended a visual angle of $5.6\circ$ wide $\times 5.6^{\circ}$ high.

In the recognition phase, a fixation cross and a preview of the two colors for the subject to choose from appeared for 750 ms followed by the test probe presentation for 750 ms. The two colors in the preview were set to two colors selected from six- and two-color conditions depending on the block. If the test probe was in the preceding study list, its corresponding color of the surrounding frame was given in the preview. After the presentation of the test probe, a long fixation cross for 1500 ms was given. Then the subjects were given two consecutive questions with an ISI of 625 ± 125 ms following each response. The first question had three options: two colors (given as solid color squares) presented in the preview and a new judgment. If the subjects recognized the probe as the study item and responded with the color they remembered as the source information of the probe in the first question, they were then asked to give a modified R-K judgment in the second question. There were three options in the R-K judgment question: remember color (give as RC), remember other (given as RO), and familiar (given as F). Subjects were instructed to answer with RC if they remembered the source information of the probe, RO if they remembered information other than the source information, and F if they could not recollect

any detail of the probe but it looked familiar. If a new judgment was made in the first question, the subjects were then asked how confident they were about the test probe being a new item: sure (given a S) or maybe (given as M). See Figure 2.1 (b) for an illustration of the study and test tasks in the experiment. The visual angle of each test probe image was 4.3° wide $\times$4.3° high. The pseudo-random response keys and mapping method from Experiment one was used except the keys for the colors and the key for the new judgment were assigned to different hands and the outer most keys (zx or ./).

EEG was recorded under the same system and setting for Experiment 1.

2.3 Experiment 3

2.3.1 Participants

The subjects were University of Colorado undergraduate students ranging in age from 18 to 29 years old (mean age = 20.6) who volunteered for course credit (21 male, 17 female) or paid participation ($15 per hour). All subjects were right-handed native English speakers and had normal or corrected-to-normal vision.

2.3.2 Experimental Paradigm and EEG Acquisition

Experiment 3 consisted of two separate sessions conducted on separate days. There were four study lists in each session, and two lists followed the location source information paradigm (as in Experiment 1) and the other two followed the color source information paradigm (as in Experiment 2). Two frame colors (blue and yellow) were selected for the color condition in order to match the number of location and color conditions across study lists. For the first session, half of the subjects received the location condition in odd list numbers and the color condition in the even list numbers. The other half of the subjects received the location and color condition in the even list numbers and odd list numbers, respectively. The second session used the opposite order. The stimuli for this experiment were taken from the stimuli in the two previous experiments.

In the study phase, for both source conditions, a source indicator frame, which was a white frame on either left or right side of the screen for location source condition and was a blue or yellow frame for color source condition, appeared on top of the visual Gaussian noise image prior to the presentation of each study item for 500 ms. The study item was then displayed inside the source indicator frame for 2000 ms followed by a ISI of 1125 ± 125 ms.

The timing of the recognition phase was identical to the previous experiments. However, there were a couple of changes in the procedure in order to match two conditions. There was no longer a color preview prior to probe presentation during the color condition lists. Also, letters B and Y were used to represent blue and yellow instead of solid color squares. Finally, both of the source responses (B and Y/L and R) were assigned to one hand and the new response was assigned to the other hand. The key assignments were again pseudo-random and counterbalanced across subjects.

EEG was recorded with the same equipment as in the previous experiments except with 500 Hz sampling frequency and without using a 0.1 Hz hardware high-pass filter.

2.4 Preprocessing

Four subjects, two subjects, and twelve subjects in Experiment 1, 2, and 3, respectively, were excluded from further analysis due to the lack of familiar responses (2 for Exp 1; 4 for Exp 3), low memory retrieval accuracy (1 for Exp1; 2 for Exp 2), less than 15 artifact-free trials (1 for Exp 1; 3 for Exp 3), or incomplete sessions (5 for Exp 3). Electroencephalography from the recognition phase of each experiment were average-referenced and epoched. Data from Experiment 3 were down-sampled to 250 Hz to match the sampling rate of Experiment 1 and 2. Each epoch was filtered between 0.1 and 50 Hz using a 40 tap FIR filter and baseline corrected by the offset from -200 ms to 0 ms.

2.5 Chapter Acknowledgements

Chapter 2, in part, is a reprint of the material as it appears in Frontiers in Human Neuroscience, 2018. Eunho Noh; Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. "Single-trial EEG analysis predicts memory retreival and reveals source-dependent differences", Frontiers in Human Neuroscience, 12, 258, 2018. The dissertation author was the primary co-author of this paper.

Chapter 3

Predicts Memory Retrieval within Subjects

3.1 Introduction

Previous recognition memory studies have used electroencephalography (EEG) to identify neural substrates of recognition memory. The 'parietal old/new effect' is a positive- going event-related potential (ERP) typically observed in the parietal electrodes between 500-800 ms and typically left lateralized. It shows greater amplitude for the correctly recognized old (hits) compared to the new (correct rejections) test items. It has been found that this effect correlates with the amount of information retrieved from the study episode (Wilding & Rugg, 1996; Wilding, 2000; Curran, 2000; Rugg & Curran, 2007), hence it is understood as a neural correlate of recollection. The 'frontal old/new effect' (or the FN400) is a frontally distributed and negative-going ERP which peaks earlier around 400 ms. The FN400 is interpreted as a neural correlate of familiarity since it shows a more negative peak for less familiar items while it typically does not vary for different amounts of recollected context information (Curran, 2000; Rugg & Curran, 2007). Pattern classification methods have been recently applied to EEG data to reveal novel findings during encoding of episodic memory (Jafarpour, Fuentemilla, Horner, Penny, & Duzel, 2014; Noh, Herzmann, Curran, & de Sa, 2014; Anderson, Zhang, Borst, & Walsh, 2016; Ratcliff, Sederberg, Smith, & Childers, 2016). In Noh et al. (2014), the classifier was used as a discriminative dimensionality reduction method to project the high-dimensional

EEG data onto a discriminative space. These projections revealed neural correlates of levels of encoding in the pre- and during-stimulus periods of the study phase. This multivariate analysis directly controls for the multiple comparison problem (MCP) by effectively reducing the number of test variables. A major advantage of this approach is that it is possible to compare the brain activity across conditions even when the trial count is low, provided that a sufficient number of classifier training trials are used to establish the initial hyperplane(s) (Noh & de Sa, 2014). Hence conditions that divide subtle behavioral differences can be readily compared. In ERP studies, these data are usually ignored or combined with other conditions to acquire reasonable ERPs for analysis. This may result in losing the ability to reveal the neural mechanisms underlying subtle behavioral differences.

In this chapter, we aim to create classifiers to discriminate between the correctly identified old/new trials during the recognition phase of episodic memory experiments on a single trial basis. We also utilize pattern classifiers as multivariate analysis tools to analyze the brain activity during retrieval of recognition memory using the time domain information of the EEG data. The EEG data were collected from 3 separate visual memory task experiments with extrinsic source information. Two types of source information were considered in these experiments. Spatial information (the location of the item) was of interest in Experiment 1 and extrinsic color information (the color of an external frame) was of interest in Experiment 2. In Experiment 3, both source types were considered. Data collected from these experiments were used to conduct multivariate analysis via pattern classifiers. The data used were previously collected by Mollison and Curran (2012), and more details were provided in Chapter 2. Mollison and Curran (2012) found that even familiar judgements were associated with above chance source judgements and that the FN400 distinguished between the source-correct and source-incorrect responses only for the location-source information but not the box-color source information. In this chapter, we specifically train separate classifiers to extract information related to item memory (without correct source memory) and source memory (for correctly remembered items) to observe any source-dependent differences that the classifiers extract between the experiments with different

source types.

The average projection values (or classifier scores) of the different source retrieval conditions and different subjective rating conditions are also compared to reveal the relationship between the different conditions and memory retrieval strength. Furthermore, data from the error conditions (incorrectly identified new trials, incorrectly rejected old trials) are projected onto the discriminative vector characterized by the different classifiers. The average projection values of these error trials are compared to those given by the other conditions and across the different projection directions.

3.2 Methods

The datasets, experimental paradigms for data collection, and the data preprocessing were the same as described in Chapter 2. In this section, we will only focus on the methods after EEG data preprocessing.

3.2.1 Classification Problems

Classification analysis was conducted separately on Experiment 1, Experiment 2, location source blocks from Experiment 3 (denoted as Experiment 3-location or Exp 3-loc), and color source blocks from Experiment 3 (denoted as Experiment 3-color or Exp 3-col). The data from Experiment 3 were divided into the different source conditions in order to reveal any potential differences between the location and color conditions that may correspond to ERP differences observed in Mollison and Curran (2012).

The classifiers were trained to find the projection function onto the vector perpendicular to the decision boundary (we sometimes refer to these vectors as planes) which is characterized by the choice of the training conditions. The behavioral conditions corresponding to correct item retrieval (SC and SI) and correct item rejection (CR) were selected for training. As a result, three different two-class binary classifiers (SC-CR, SI-CR, and SC-SI) with probability outputs ($0 \leq p \leq 1$) were trained to discriminate between pairs of behavioral conditions. These

probability outputs given by the classifiers are denoted as classifier scores in this paper. The classifiers were trained on each individual subject and only the subjects with a minimum of 25 trials for each of the 2 trained conditions (SC, SI, and CR) were included in the analysis. For each classification problem, the classifier scores were also computed for the trials which were not included in the training procedure (non-training trials).

- SC-CR classifier

  The SC-CR classifier (trained to discriminate between SC and CR) was expected to find the projection which maximizes the difference in the amount of information retrieved from the study episode.

- SI-CR classifier

  This classifier (trained to discriminate between SI and CR) was designed to discriminate between correctly retrieved old items (with incorrect source judgments) and the correctly rejected new items

- SC-SI classifier

  The SC-SI classifier (trained to discriminate between SC and SI) was built to distinguish the correctly retrieved old items with correct source judgments from those with incorrect source judgments. Hence the classifier would extract the information on source memory retrieval.

3.2.2 Classification

The spatio-temporal structure of the ERPs was extracted based on previous findings on the old/new effect. Six channel groups were selected for evaluation (LAS, RAS, CM, LPS, RPS, and PM) as given in Figure 2.2. The average voltage for each channel group was computed and the data between 300 and 800 ms after test item presentation were extracted to take advantage of the ERP effects related to memory retrieval. The dimensionality of these subsequences were reduced to 5 by averaging over 100 ms length non-overlapping windows. The features from all

six channel groups were concatenated to build a 30-dimensional feature vector for each trial. A binary classifier using linear discriminant analysis (LDA) with automatic shrinkage (Ledoit & Wolf, 2004; Schäfer & Strimmer, 2005) was trained to classify these feature vectors (Lotte, Congedo, Lécuyer, Lamarche, & Arnaldi, 2007; Blankertz, Lemm, Treder, Haufe, & Müller, 2011). In order to avoid any overfitting to the training data, the projections for the training conditions were computed using leave-two-out (one from each class) cross-validation. In order to train with balanced classes, trials from the majority class were randomly discarded (from training) to have equal numbers of trials in each class. These trials however were still used for evaluation of the classifier (using a classifier trained on all the selected balanced training data). The data from the remaining conditions (e.g. Misses and False Alarms) were not used to evaluate the classifier, but were still projected onto the discriminative vector (learned from the entire balanced training set) for interpretative analysis.

3.2.3 Statistical Methods

The average classifier scores (for a given classification problem) across all subjects were compared across different behavioral conditions (SC, SI, CR, M, and FA). The classifier score is a projection of the high-dimensional EEG data onto a 1-dimensional vector which is representative of the given classification problem. Paired t-tests were conducted on the trial-by-trial classifier scores separately for the four available datasets to compare the classifier scores of the different retrieval/subjective rating conditions. A comparison was considered to be significant only when all four separate datasets gave p-values below 0.05 for the conditions of interest.

It is advantageous to also visualize the EEG features utilized by the classifiers for interpreting any effects identified from the multivariate analysis using the pattern classifiers. This was done by analyzing the classifier activation patterns representing which channel, time pairs were important for classification (Haufe et al., 2014). The activation pattern A of the LDA

classifier (Haufe et al., 2014) could be expressed as,

$$A \propto \Sigma_x W = \hat{\mu}_1 - \hat{\mu}_0$$

where Σ_x was the covariance of the training data, W was the discriminant vector of the LDA classifier, and $\hat{\mu}$ was the estimated mean of each class. For each source type, the 30-dimensional classifier activation pattern vector for each subject was normalized to have length 1.

In order to identify features consistent across subjects, a cluster-based method for correction for multiple comparisons was used (Maris & Oostenveld, 2007). In this method, first each spatiotemporal pixel significantly different from zero ($p<0.05$) was identified. Then the t-statistic of all significant flagged neighboring pixels with the same sign was summed and the maximum absolute value over all clusters taken. This value is compared to the distribution of max absolute cluster values obtained from a permutation distribution resulting from 10,000 random permutations of class labels for each subject. Temporal neighbors were temporally adjacent time windows. Spatial groups were considered neighbors if they contained adjacent electrodes from the cap layout (see Figure 2.2). Using this rule, LAS, CM, and RAS were all mutual neighbors; CM was also neighbors with LPS and RPS; LPS and RPS are also neighbors with PM.

3.3 Results

3.3.1 Classifier Performance

Performance of the SC-CR classifier was computed based on classification of the SC and CR trials (SC-RS, SC-RO, SC-F, CR-SN, CR-MN). The significance of the performance of a classifier (whether it performs significantly over chance) was evaluated based on the number of test trials used for classification. The 95% confidence interval for the obtained accuracy was calculated using Wald intervals with small sample size adjustments (Agresti & Caffo, 2000) for each subject. Classification results were considered to be significantly over chance only

when the interval did not include 50%. Results are given in Table 3.1, 3.2, and 3.3. The overall classification accuracy for Experiment 1 (SC-CR) was 62% with 18 of 25 subjects having individual accuracies significantly over chance. When restricted to subjects with at least 50 trials in each class, the performance is somewhat better. The overall classification accuracy for Experiment 2 (SC-CR) was 59% with 17 of 28 subjects having individual accuracies significantly over chance. Experiment 3-loc (SC-CR) had an average accuracy of 57% and Experiment 3-col (SC-CR) had an average accuracy of 56%.

Table 3.1. Classification results for Experiment 1.

Sub. ID	**SC-CR**	**SI-CR**	**SC-SI**
102	0.5538	*0.4702*	*0.4693*
103	*0.4857*	-	-
104	**0.6875**	0.5572	**0.6540**
106	**0.6720**	0.5455	0.5358
108	**0.6550**	0.4891	0.4949
109	0.5593	-	-
110	**0.5947**	*0.4712*	*0.5000*
112	*0.4953*	-	-
113	**0.6667**	*0.3280*	***0.6271***
114	**0.6746**	0.5575	**0.6442**
115	**0.5741**	0.5269	0.5517
116	0.5251	0.5183	0.4839
117	**0.5944**	*0.5148*	*0.5025*
118	**0.6500**	*0.4762*	*0.5398*
119	**0.6154**	**0.5756**	0.5000
120	**0.6172**	*0.5108*	***0.6090***
121	**0.5977**	0.5057	0.5356
122	0.5255	***0.6224***	*0.5420*
123	**0.6585**	*0.5603*	*0.5368*
124	*0.5649*	*0.5577*	*0.6104*
125	**0.6518**	0.5571	0.4373
126	**0.6955**	*0.4419*	***0.5981***
127	**0.6048**	*0.5530*	*0.5194*
128	**0.7474**	*0.4633*	*0.5302*
129	**0.5542**	*0.4328*	*0.4714*
Overall	0.6231	0.5090	0.5383
Overall with 50 trials/class cutoff	0.6290	0.5368	0.5397

Overall accuracies given in the penultimate row are the accuracies over all trials from the relevant classes for subjects with 25 or more trials per class. Overall accuracies in the last row are computed over all trials from relevant classes for subjects with 50 or more trials per class. Bolded entries are significantly better than chance ($p<.05$). Results from subjectrs with less than 50 trials per condition are italicized

Figure 3.1 gives the ROC (receiver operating characteristic) curves for choosing different thresholds (between 0 and 1) to make decisions between classes 1 and 2 for all 3 classification problems. Table 3.4 gives the area under these ROC curves. All results were above 0.5 however

Table 3.2. Classification results for Experiment 2.

Sub. ID	SC-CR	SI-CR	SC-SI
201	*0.5113*	*0.5581*	0.5298
202	***0.4857***	*0.5915*	*0.4713*
203	*0.5444*	*0.4805*	*0.4773*
204	**0.6018**	*0.4800*	***0.5787***
205	**0.5766**	0.4303	0.5126
206	0.5524	*0.5816*	*0.4563*
207	**0.6204**	*0.4177*	*0.4912*
208	*0.6349*	0.4643	**0.5701**
209	***0.6222***	*0.5714*	0.5344
210	**0.6189**	0.5356	0.5193
211	**0.6000**	0.4851	0.5193
212	*0.4964*	*0.5146*	0.5723
213	**0.5942**	**0.5922**	0.5071
214	**0.6553**	0.5659	0.5211
215	**0.6728**	**0.6554**	0.5270
216	0.5475	0.5315	0.5571
217	**0.5761**	0.5484	0.5200
219	0.5506	*0.4859*	*0.4757*
220	**0.6174**	**0.6171**	0.5436
221	**0.5737**	0.4826	0.5345
222	0.5504	0.4758	0.5780
223	0.5500	0.4943	0.4775
224	**0.5789**	**0.5822**	**0.6053**
225	0.5538	**0.5952**	0.5431
227	**0.6757**	**0.6000**	0.5410
228	*0.5533*	*0.5784*	0.5287
229	0.4588	0.4262	0.4456
230	**0.5771**	**0.5805**	0.5171
Overall	0.5904	0.5383	0.5263
Overall with 50 trials/class cutoff	0.5945	0.5416	0.5311

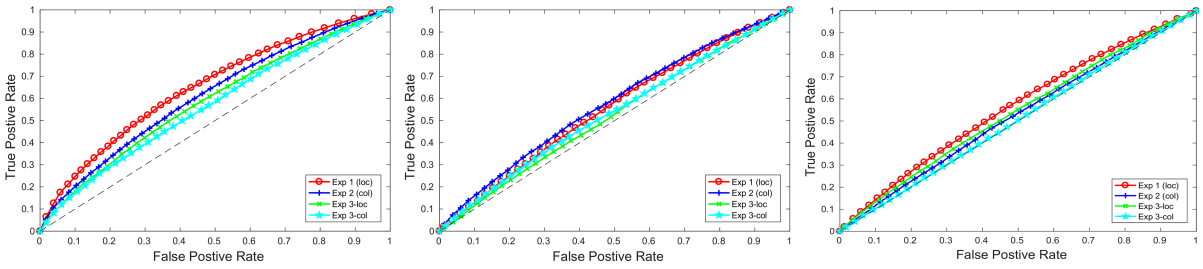
Overall accuracies given in the penultimate row are the accuracies over all trials from the relevant classes for subjects with 25 or more trials per class. Overall accuracies in the last row are computed over all trials from relevant classes for subjects with 50 or more trials per class. Bolded entries are significantly better than chance ($p<.05$). Results from subjectrs with less than 50 trials per condition are italicized

there was a variability in performance across the different classification problems. The SC-CR classifiers showed the highest performance on all 4 datasets. It was also found that the datasets with recordings from multiple days (Exp 3-loc and Exp 3-col) showed a slight decrease in performance compared to the single session datasets. The SC-SI classification performs better for the location source datasets relative to the color source datsets in contrast with the SI-CR classifiers.

Table 3.3. Classification results for Experiment 3.

Sub. ID	SC-CR (loc)	SC-CR (col)	SI-CR (loc)	SI-CR (col)	SC-SI (loc)	SC-SI (col)
310	**0.5665**	0.4833	*0.4538*	0.5000	*0.5289*	0.5208
312	**0.6422**	**0.6257**	*0.5380*	0.5451	*0.4591*	0.5208
313	**0.5700**	**0.5766**	-	0.5509	-	0.5498
315	0.5127	**0.5949**	0.5074	0.5505	0.5181	0.5064
317	**0.6039**	**0.6101**	0.5153	0.5000	0.5019	0.4704
318	*0.5327*	*0.5469*	*0.6026*	*0.5185*	*0.5315*	*0.5041*
319	0.5550	*0.4706*	0.4581	*0.5399*	0.5444	0.4681
321	0.5613	**0.6291**	**0.5985**	0.5369	0.5076	**0.5647**
322	*0.5252*	*0.4818*	*0.4811*	*0.5076*	0.5170	0.4696
323	0.5615	0.5291	*0.4538*	0.5505	*0.5519*	0.5528
324	0.5244	0.5393	0.5407	0.5056	0.5430	0.5358
326	**0.6181**	**0.5700**	0.4928	0.4962	0.5516	0.4697
327	**0.6111**	0.5022	-	0.5586	-	0.5054
328	0.5120	0.4971	*0.5515*	0.5200	*0.5426*	0.5137
330	**0.6348**	**0.5638**	0.4462	0.4908	**0.6035**	0.4967
332	**0.5710**	**0.5698**	-	0.4632	-	0.4615
333	0.5593	0.5445	-	0.5724	-	0.5364
334	*0.5370*	0.4563	*0.5149*	**0.5984**	0.5195	0.4978
335	0.4963	**0.5714**	-	*0.5000*	-	*0.4701*
336	**0.6262**	**0.6313**	*0.4545*	0.4527	***0.5685***	0.4963
337	**0.5638**	0.5481	0.5546	0.5037	0.5547	0.5000
340	*0.4875*	**0.6138**	*0.3663*	0.4965	0.4785	0.5134
342	**0.5980**	0.5233	*0.4710*	0.5425	*0.4651*	0.4870
343	**0.6667**	*0.5179*	0.5652	***0.5870***	0.5059	0.4805
344	**0.5682**	0.5158	0.4930	0.4667	0.5081	0.4727
345	0.5101	0.4688	*0.5429*	0.5058	*0.5587*	0.5053
Overall	0.5736	0.5553	0.5049	0.5182	0.5261	0.5024
Overall with 50 trials/class cutoff	0.5789	0.5610	0.5153	0.5159	0.5286	0.5031

Overall accuracies given in the penultimate row are the accuracies over all trials from the relevant classes for subjects with 25 or more trials per class. Overall accuracies in the last row are computed over all trials from relevant classes for subjects with 50 or more trials per class. Bolded entries are significantly better than chance ($p<.05$). Results from subjectrs with less than 50 trials per condition are italicized



(a) SC vs CR classification (a) SI vs CR classification (a) SC vs SI classification

Figure 3.1. The ROC curves for the three different classification problems (A: SC vs. CR; B: SI vs. CR; and C: SC vs. SI) are given separately for the four individual datasets.

Table 3.4. Classification results for Experiment 1 to 3.

	SC-CR	**SI-CR**	**SC-SI**
Experiment 1 (loc)	0.6555	0.5586	0.5434
Experiment 2 (col)	0.6160	0.5779	0.5357
Experiment 3 (loc)	0.5916	0.5264	0.5375
Experiment 3 (col)	0.5726	0.5376	0.5108
Average	0.6089	0.5501	0.5319

Every AUC was computed using the projections of the trials, by leave-two-out training, in the selected balanced training data.

3.3.2 Analysis of the Classifier Scores

The projection weights for a given classification problem can be used to project the EEG data onto a discriminative vector. In this chapter, these projection values are denoted as the classifier scores. The relationship between the average classifier scores for the different behavioral conditions represent the characteristics of the different discriminative hyperplanes (Noh & de Sa, 2014). As described in Section 3.2.3, the representation of the EEG data on the three different discriminative vectors were compared across the different behavioral conditions. The classifier scores were computed for each classification problem and the average scores corresponding to the different behavioral conditions were compared. The results were compared across the four datasets and effects with $p < 0.05$ consistently across the different datasets were considered to be meaningful (the individual comparison results are given in Table 3.4).

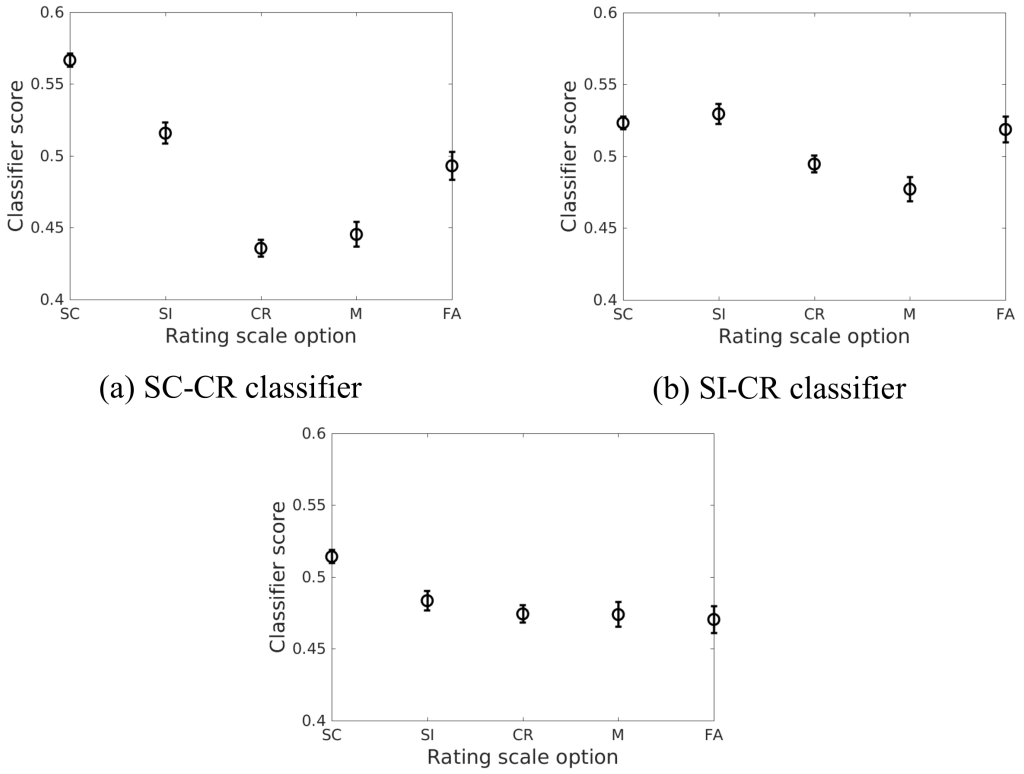
Table 3.4. Comparison results between the classifier scores for the SC-CR classifier.

	SC vs. CR	SC vs. SI	SC vs. M	SC vs. FA	CR vs. SI	CR vs. M	CR vs. FA	SI vs. M	SI vs. FA	M vs. FA
Exp 1 (loc)	5.79E-110	4.11E-11	6.45E-47	9.77E-8	5.22E-15	0.2436	1.08E-6	5.56E-12	0.2108	5.07E-8
Exp 2 (col)	5.53E-49	2.46E-3	1.93E-14	9.28E-4	3.29E-27	0.1460	2.93E-7	6.62E-7	0.5146	2.22E-3
Exp 3-loc	6.86E-34	5.46E-3	1.36E-25	4.43E-10	3.89E-6	0.8663	4.30E-3	1.34E-7	0.0660	1.29E-3
Exp 3-col	3.63E-27	9.87E-5	2.25E-9	2.27E-7	7.60E-11	0.7677	0.2395	1.85E-4	1.12E-4	0.1455

Paired t-tests between all possible pairs are given with their corresponding uncorrected p-values.

The correct item memory conditions (SC, SI, and CR) showed similar patterns across the different projections where SC trials gave the highest scores and the CR trials showed the lowest scores. However, the relative distance between the three conditions varied across the different

discriminative vectors. It was found that the SI condition was mapped closer to the CR condition on the SC-SI plane (see Figure 3.2C) while it was mapped closer to the SC condition on the SI-CR plane (see Figure 3.2B). It was also found that the difference between the SI and CR trials were only significant ($p < 0.05$ for all four datasets) on the SC-CR and SI-CR planes (see Figure 3.2).



(a) SC-CR classifier

(b) SI-CR classifier

(c) SC-SI classifier

Figure 3.2. The average of the estimated means and the approximate 95% confidence intervals of the classifier scores across the four datasets [Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col] for the five behavioral conditions (SC, SI, CR, M, and FA) for the three different classification problems (A: SC vs. CR; B: SI vs. CR; and C: SC vs. SI).

The relative mapping of the error conditions (M and FA) with respect to the correctly retrieved/rejected conditions (SC, SI, and CR) gave different patterns for the different projection directions. Interestingly, the source correct (SC) trials and false alarms (FA) were mapped to significantly different values on the SC-CR and SC-SI plane but not on the SI-CR plane (see

Table 3.5. The uncorrected pairwise comparison results for the five behavioral conditions across the four datasets [Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col].

	SC vs. CR	SC vs. SI	SC vs. M	SC vs. FA	CR vs. SI	CR vs. M	CR vs. FA	SI vs. M	SI vs. FA	M vs. FA
SC-CR	2.35E-193	1.75E-28	6.41E-83	4.86E-26	9.18E-49	0.0556	8.80E-14	2.07E-25	0.0504	6.10E-14
SI-CR	3.46E-10	0.1210	8.04E-10	0.5285	1.44E-11	0.0238	9.54E-5	9.27E-14	0.3268	3.38E-11
SC-SI	1.52E-23	6.84E-6	8.67E-11	1.36E-10	4.31E-3	0.1911	9.83E-3	1.45E-3	0.3237	

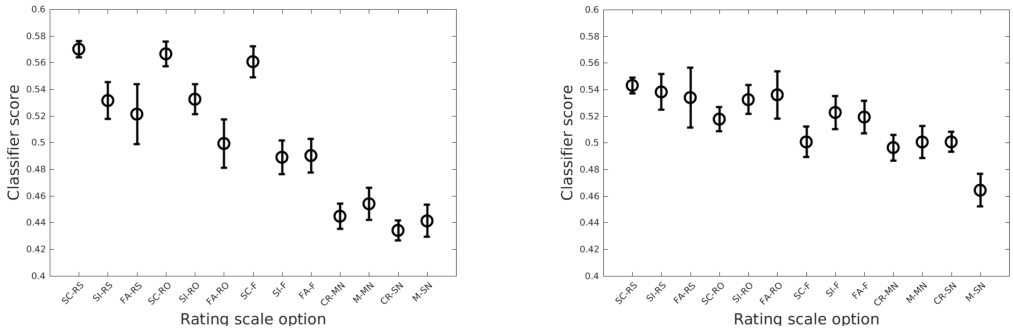
The results for the different projections are given in separate rows

Table 3.5). In contrast, the misses (M) gave values significantly lower than the two correct item retrieval conditions (SC and SI) when mapped onto the SC-CR and SI-CR plane.

A similar analysis was conducted considering the different subjective ratings given to the correct item retrieval/rejection trials (SC, SI, and CR). These responses consisted of remember source (RS), remember other (RO), and familiar (F) for the SC/SI conditions and sure (SN denoting sure new) and maybe (MN denoting maybe new) for the CR condition. The error conditions (FA and M) can be similarly projected. While the classifiers generally gave a monotonic decrease in classifier scores from the RS to SN conditions, there were interesting interactions with the memory retrieval conditions as illustrated in Figure 3.3.

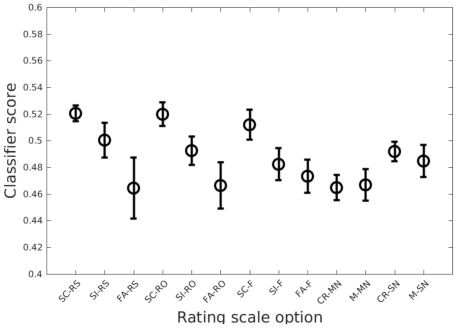
3.3.3 Classifier Activation Patterns

The activation patterns which represent the features used by the classifiers (or the characteristics of the projection weights) were compared across the three different classification problems (see Figure 7). The activation patterns were computed for each subject and the average activation patterns were computed by averaging the values across all four datasets. A t-test was conducted on each of the features to illustrate which features showed similar effects across the different subjects. Cluster based analysis (Maris & Oostenveld, 2007) was then used to control for multiple comparisons. This revealed features with values significantly above/below zero across all the subjects available for analysis. The activation patterns are given as a 2-dimensional matrix with its corresponding channel groups and time segments (the times give the center of the interval) in Figure 3.4 and the most significant clusters (with significance values) are shown in Figure 3.5.



(a) SC-CR classifier

(b) SI-CR classifier



(c) SC-SI classifier

Figure 3.3. The average of the estimated means and the approximate 95% confidence intervals of the classifier scores across the four datasets (Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col) when considering the breakdown by subjective ratings (RS, RO, F, MN, and SN) for the three different classification problems (A: SC vs. CR; B: SI vs. CR; C: SC vs. SI).

The SC-CR classifier utilized temporal features from 300 to 800 ms. The SI-CR classifier only showed consistent patterns between 300 and 700 ms and the SC-SI classifier showed consistent patterns between 400 and 800 ms for the two tasks with spatially presented contextual information (1 and 3-loc). In the two tasks with colored frames as context, there is not a strong activation pattern consistency across subjects for the SC-SI classifier. Interestingly the SC-SI (source memory) classifier has strong consistent activity across all spatial areas except PM when the source context is location. The SI-CR (item memory) classifiers have an early frontal activation when the source context is the colored outline, but a more parietal activation when the source context is the location.

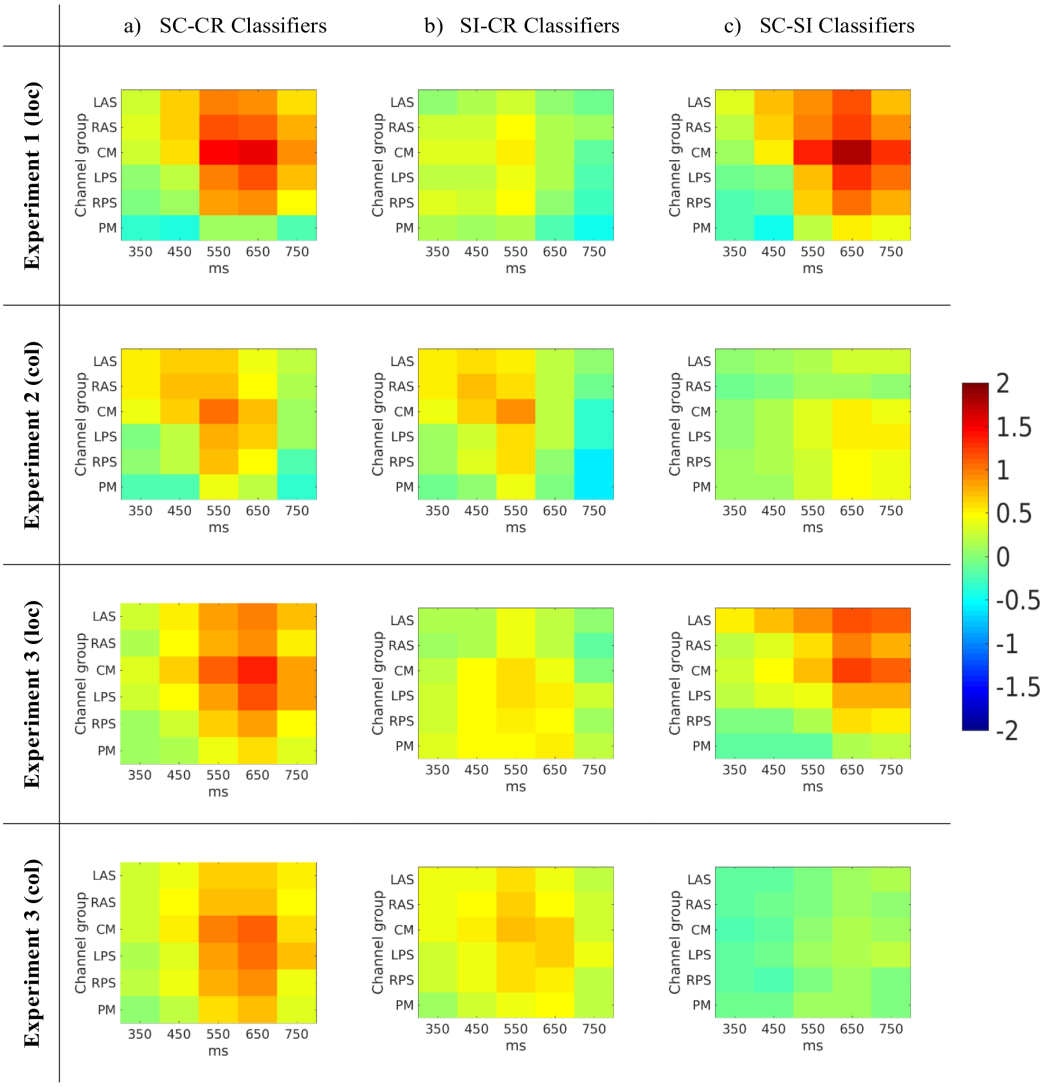


Figure 3.4. (A) The average activation patterns from the SC-CR classifiers averaged across all available subjects. (B) The average activation patterns from the SI-CR classifier averaged across all available subjects. (C) The average activation patterns from the SC-SI classifiers averaged across all available subjects. Note that the numbers on the x-axes represent the mid-point of the 100 ms window used to compute the features.

3.3.4 Classifier Scores Evolution over Time

In the activation patterns, the characteristics of projection weights in different time intervals and channel groups were shown. The classifier scores variation across time gives a

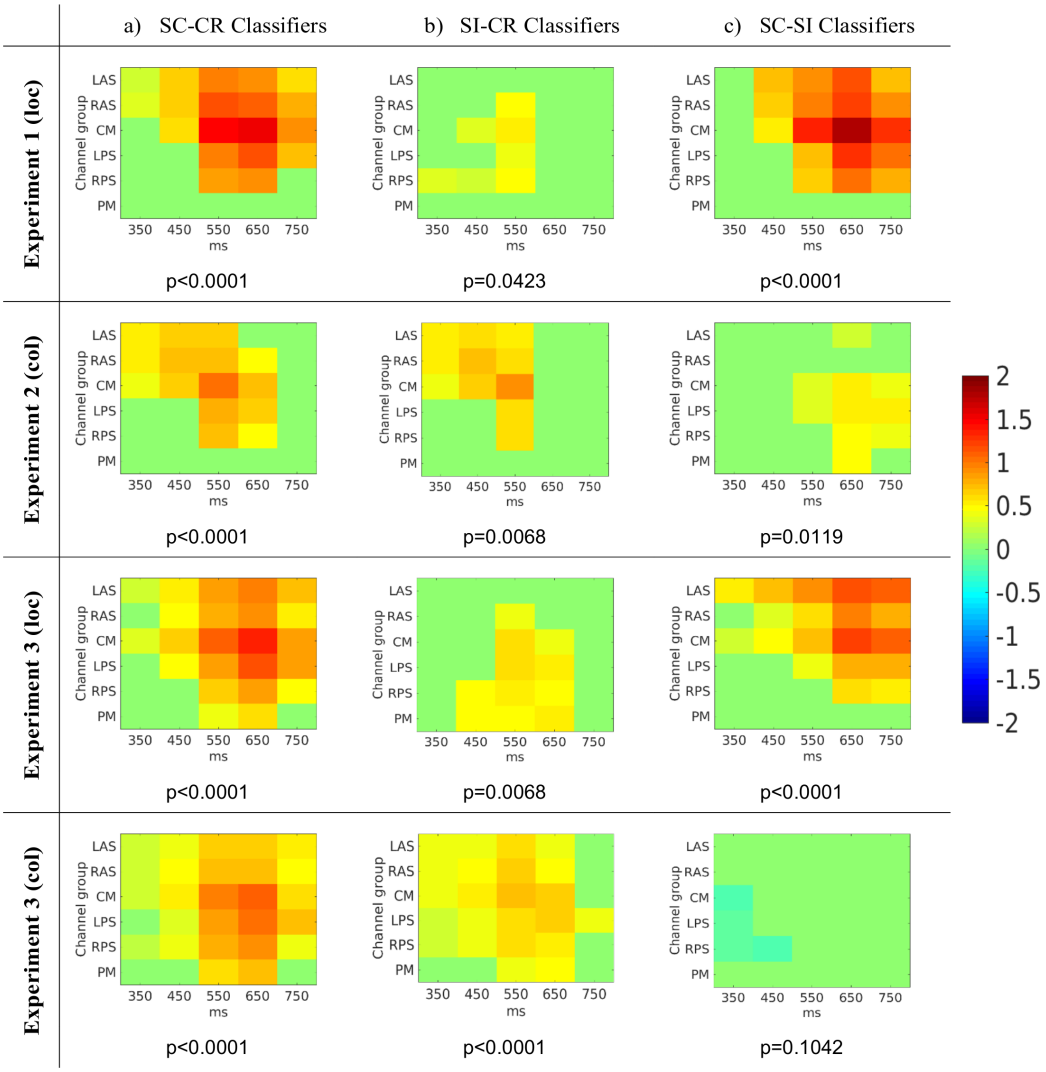


Figure 3.5. The average activation patterns from Figure 3.4 but masked by the most significant cluster (with p-value for that cluster given below) (Maris & Oostenveld, 2007) for the three different classification problems (A: SC vs. CR; B: SI vs. CR; C: SC vs. SI). Note that the numbers on the x-axes represent the mid-point of the 100 ms window used to compute the features.

clear insight about the evolution of the separation of classes over time. To obtain the scores only under the operation with weights between 300 and 400 ms in activation patterns, the grouped EEG data after 400 ms were set to zero, and the remaining computations remained the same. In

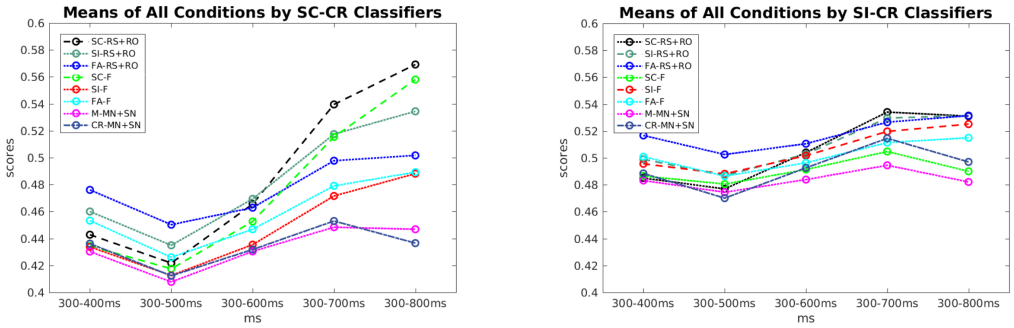
brief, the data were set to zero after the considered intervals and the trained classifier was used to get the classifier scores.

Figure 3.6A shows that the scores of SC and CR trials start to be discriminable around 500–600 ms and separate further afterwards. Figuree 3.6B, shows that with the SI-CR classifier, scores of SI and CR trials also start to separate around 500–600 ms. As for the SC-SI classifier in Figuree 3.6C, the scores of SC trials become more separable from the scores of the SI trials after about 700 ms. Note that while the activation patterns for the SI-CR classifier show not much significant activation that is consistent between subjects after 600 ms, the classifier scores continue to separate, indicating that the activation patterns causing this separation are less consistent between subjects.

3.4 Discussion

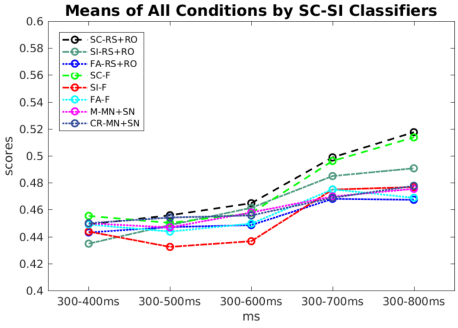
The results show that it is possible to predict successfully identified old vs. new items based on single-trial scalp EEG activity recorded during the retrieval episode. The prediction rate was higher for the location-source datasets and the average accuracy of the single-session datasets was higher compared to the multi-session datasets. The non-stationarity of the data between the two sessions (due to electrode position changes, impedance changes, or changes in brain-state) likely contributes to the drop in classification performance (Krauledat, Schröder, Blankertz, & Müller, 2007). Our analysis was restricted to time domain signals from specific channel groups known to be involved in frontal and parietal old/new effects. It is possible that accuracy could be increased by using frequency domain information from multiple electrode location and frequency bands (see for examples in Hammon and de Sa (2007); Hammon, Makeig, Poizner, Todorov, and De Sa (2007); Velu and de Sa (2013); Noh et al. (2014); Mousavi, Koerner, Zhang, Noh, and de Sa (2017)).

The current analysis found that the projections of the temporal information from the EEG data onto different hyperplanes show different patterns. This was evident in the relationship



(a) SC-CR classifier

(b) SI-CR classifier



(c) SC-SI classifier

Figure 3.6. (A) The scores of all conditions across time by SC-CR classifiers. (B) The scores of all conditions across time by SI-CR classifiers. (C) The scores of all conditions across time by SC-SI classifiers.

between the behavioral conditions of interest. We focused on the patterns which were consistent across multiple subjects and multiple datasets to compare across the different classifiers. These results suggested that the classifier may be exploiting features which are more informative for discriminating between the two behavioral conditions selected for training. It was found that the SC-SI classifier performance on the two color source datasets (Experiment 2 and the color blocks from Experiment 3) was lower compared to the location datasets (Experiment 1 and the location blocks from Experiment 3). The activation patterns for the SC-SI classifier were also not significantly consistent across subjects for the color outline source. In Mollison and Curran (2012), it was found that accurate/inaccurate judgments to the familiar responses were affected

by source type where the SC trials with familiar ratings and SI trials with familiar ratings were significantly different only for the location-source datasets when comparison was conducted on a ROI centered at FCz. This suggests that the temporal information in the EEG signal may be less separable between SC and SI trials for the color datasets compared to the location datasets resulting in a lower classification performance.

The relationship between the correctly remembered conditions (where the classifier scores showed CR $<$ SI $<$ SC on all three discriminative vectors) suggests that these classifier scores may reflect the amount of information retrieved from the study episode. The difference in the amount of information retrieved from the study episode is maximal between conditions SC (when the correct item is retrieved from the study phase with the appropriate source information) and CR (when no information is retrieved from the study phase) which may be why the SC-CR classifier outperformed the other two classifiers. The drop in classifier performance for the SC-SI and SI-CR classifiers compared to the SC-CR classifier may be due to this innate relationship between the 3 behavioral conditions used for classifier training. The SI-CR classifier would primarily be able to utilize information related to differences between item retrieval vs. correct rejection to distinguish between the two classes. On the contrary, the SC-SI classifier would only be able to utilize information related to source memory differences between correct source retrieval vs. incorrect source retrieval in order to distinguish between the SC and SI conditions.

The activation patterns (see Figure 3.5) indicated that the classifiers used features mostly around 400 to 800 ms and gave these features higher weights. The spatiotemporal distribution of predictive features associated with the (SI - CR) classifiers (early and more frontal) were somewhat consistent with the timing and location of the FN400 only in the color-source experiments [2 and 3(col)]. Likewise the spatiotemporal distribution of predictive features associated with the (SC-SI) classifiers (later and more parietal) were somewhat consistent with the timing and location of the parietal ERP old/new effect only in the color-source experiments. In the location-source experiments, the (SC-SI) classifier had significant contributions from both early ($<$ 500 ms) and late (500–800) time periods and frontal and parietal locations. This suggests that

while the SI-CR classifier may be representative of the early frontal old/new effect and the SC-SI classifier representative of the later parietal old/new effect when color is the source information, the mapping is not as appropriate when location is the source information. This is consistent with Mollison and Curran (2012) observations suggesting that familiarity contributes to source recognition for location more so than for color. The activation patterns corresponding to the SC-CR classifier took advantage of the features across all time periods (see Figure 3.5) which most likely resulted in the largest distinction between the SC and CR condition.

Additionally, the multivariate classification approach showed that trial-by-trial variation in EEG corresponding to these ERP components are predictive of subjects' behavioral responses, which is consistent with the hypothesis that the underlying processes are influencing memory judgments. One previous study has similarly used logistic regression to predict performance on a city-size comparison task from single-trial EEG data corresponding to the FN400 (Rosburg, Mecklinger, & Frings, 2011). Their results showed that the relative familiarity of two cities, as indexed by single-trial FN400 measures, predicted which of the cities subjects judged as being more populous. Taken together with the current results, these classification approaches are important for establishing that EEG patterns which have been related to familiarity and recollection in ERP averages, can be shown to predict behavior on individual trials in both standard memory tasks as well as a decision making task that is influenced by memory. Overall, this strengthens the hypothesized links between these EEG patterns and behaviorally relevant memory processes.

The ERP studies of recognition memory often exclude error trials from analyses because of insufficient trials for stable ERPs in these conditions. In their original study, Mollison and Curran (2012) excluded subjects with less than 15 artifact-free trials/condition/subject and 24% of subjects would have been excluded if errors were included in the analyses. One approach for increasing the false alarm rate has been to use lures that are similar to studied items (e.g., Curran (2000); Curran and Cleary (2003); Nessler, Mecklinger, and Penney (2001). In these cases subjects are presumed to have a high false alarm rate because similar lures are as familiar

as studied items, and the familiarity-related FN400 responds similarly to hits and false alarms to similar lures. It is also common to hypothesize that false alarms to even non-similar lures are driven by familiarity. For example, the dual-process model (Yonelinas, 1994, 1997) of ROC curves explicitly assumes that recollection does not contribute to false alarms, which are only driven by familiarity. Few ERP studies have assessed false alarms from lures that were not similar to the studied items. If familiarity differentiates "no" (CR) and "yes" (FA) responses to new items, the FN400 should be more positive to FA trials than CR trials. Although early studies that did not clearly differentiate the FN400 reported no differences between hits and false alarms (Wilding, Doyle, & Rugg, 1995; Wilding & Rugg, 1996, 1997; Rubin, Petten, Glisky, & Newberg, 1999), two studies that specifically focused on the FN400 did observe more positive FN400s to FA than to CR trials (Finnigan, Humphreys, Dennis, & Geffen, 2002; Wolk et al., 2006). Wolk et al. (2006) included a very large number of test items, which resulted in an average of 105 FA trials/subject, but Finnigan et al. (2002) only averaged 12 trials/subject. The current multivariate analysis approach using pattern classifiers addresses this trial count issue by projecting the high dimensional EEG data onto a one-dimensional vector which is meaningful with respect to the experimental paradigm. The SI-CR classifier responded more strongly to FA trials than to CR, with FA being more similar to item hits (SC and SI), as would be expected if FA trials were driven by familiarity.

The relationship between the SC and FA conditions was particularly interesting. The difference between the two conditions were consistently larger across all four datasets on the SC-CR and SC-SI planes compared to the SI-CR plane as given in Table 3.6 (and shown in Figure 3.2). This pattern was also evident between the SI and FA conditions, however, the distances between these two conditions were closer. Hence the representations with respect to the different classification boundaries suggest that SC and FA are more similar to each other on the SI-CR (item memory) plane compared to the other two representations. In other words, false alarms (on item information) may include information related to item retrieval while they do not include much information related to source retrieval (recollection).

Table 3.6. The difference between the average classifier scores for the SC and FA conditions are given in the top three rows.

Classifier	Exp 1 (loc)	Exp 2 (col)	Exp 3-loc	Exp 3-col	Average
Difference between SC and FA					
SC-CR	0.0992	0.0630	0.0647	0.0546	0.0704
SI-CR	-0.0220	0.0099	-0.0042	0.0151	-0.0003
SC-SI	0.0832	0.0456	0.0479	0.0241	0.0502
Difference between SI fan FA					
SC-CR	0.0121	0.0353	0.0135	0.0254	0.0216
SI-CR	-0.0173	0.0258	-0.0021	0.0096	0.0040
SC-SI	0.0244	0.0171	0.0121	0.0165	0.0175

The difference between the average classifier scores for the SI and FA conditions are given in the bottom three rows. Negative values indicate FA had larger values. The table shows that FA is mapped closer to SC and SI in the SI-CR classifier than the SC-CR and SC-SI classifiers.

The other type of error, misses (M), were generally similar to CR in all three classifiers. Both of these conditions reflect low levels of familiarity and recollection that lead to "no" responses. Previous studies have found 300–500 ms FN400 or 500–800 ms parietal old/new differences between hits and misses, but not between CR trials and misses (Rugg et al., 1998; Curran & Hancock, 2007). Instead, Rugg et al. (1998) found differences between misses and CR were observed over posterior channels between 300 and 500 ms. The latter differences were interpreted as reflecting the activity of an implicit memory process because subjects were giving the same explicit "no" response to both old and new items, but the brain was still differentiating their memory status [although others dispute this definition of implicit memory (Voss & Paller, 2008)]. Because our classifiers were trained to differentiate different levels of explicit memory, it makes sense that no major differences were observed between misses and CR in any of our results. Future work could be done to further investigate any differences by specifically involving misses in the classification training [see for example in (Noh & de Sa, 2014)].

In summary, the present results showed that the classification analysis successfully extracts information related to retrieval strength from the EEG data. These results show that the classifier scores well represent the subjects' behavioral performance on source retrieval (the relationship between the SC, SI, and CR conditions in Figure 3.2) and indicate that EEG

item-memory and source-memory responses may be more spatially widespread than previously thought and differ between source-types. The results also indicate that retrieval strength as reflected in the classifier scores follows the subjects' subjective ratings (Figure 3.3 and Table 3.7). It was also found that the brain activity related to item memory/familiarity may be present during false item retrieval (FA trials) as well as during correct item retrieval (SC and SI trials).

Table 3.7. The uncorrected pairwise comparison results for the five subjective rating options across the four datasets [Exp 1 (loc), Exp 2 (col), Exp 3-loc, Exp 3-col].

	RS vs. RO	RS vs. F	RS vs. MN	RS vs. SN	RO vs. F	RO vs. MN	RO vs. SN	F vs. MN	F vs. SN	MN vs. SN
SC-CR	3.12E-3	6.83E-27	8.57E-92	8.08E-158	5.62E-13	4.19E-66	2.31E-117	3.52E-28	1.97E-47	1.77E-3

Only the trials with correct item judgments were included in this analysis.

3.5 Chapter Acknowledgements

Chapter 3, in full, includes the result section and discussion section in the following publication in Frontiers in Human Neuroscience, 2018. Eunho Noh; Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. "Single-trial EEG analysis predicts memory retrieval and reveals source-dependent differences", Frontiers in Human Neuroscience, 12, 258, 2018. The dissertation author was the primary co-author of this publication.

Chapter 4

Predicts Memory Retrieval across Subjects

4.1 Introduction

Currently, brain-computer interfaces (BCI), which allow users to interact with devices using brain signals, are facing the challenge that a classifier for one individual might not be usable by another subject due to individual differences in brain anatomy and physiology (Abdulkader, Atia, & Mostafa, 2015). For application to different users, the classifier usually requires either re-training or calibration. This problem could be solved with a subject-independent classifier trained without using the data for the new user.

In the previous chapter, we trained individual classifiers for each subject tested using leave-one-trial-out (LOTO) cross-validation for single-trial EEG classification in memory retrieval prediction. The analysis of the activation patterns across classifiers trained on different subjects suggested that there was some consistency in the classifiers for the different subjects on the same classification problem. The authors in Fazli et al. (2009); Ray et al. (2015), proposed creating subject-independent EEG-based BCIs for motor-imagery and emotional imagery tasks and demonstrated the potential of designing EEG-based leave-one-subject-out (LOSO) classifiers for these tasks. Also, in Sun et al. (2016), a universal memory classifier was trained with cross-validation with data across all subjects and was able to distinguish remembered trials from forgotten trials across subjects. As this study did not perform LOSO, it is not clear how

important the training data from the same subject was for good performance, but it did create one classifier that worked for all subjects. In another study O'Sullivan et al. (2015), data from other subjects was used to predict auditory attentional selection. These studies motivated our present approach to train a subject-independent memory classifier using LOSO cross-validation.

In this chapter, we aim to create LOSO classifiers to discriminate between correctly identified old/new trials from a new subject during the recognition phase of episodic memory experiments on a single trial basis. We utilized the temporal information between 300 and 1500 ms, longer than used in Noh, Liao, Mollison, Curran, and Sa (2018), in order to allow for the inclusion of all the above mentioned recognition processes for old/new effects. We again used pattern classifiers as multivariate analysis tools to reduce the dimensionality (Noh & de Sa, 2014) and analyze the brain activity during the recognition phase in memory experiments using the spatio-temporal information of the EEG data.

4.2 Methods

While the datasets, experimental paradigms for data collection, and the data preprocessing remained the same as the counterparts in Chapter 2, the time window investigated was longer than the window in Chapter 3. In order to include the temporal features related to LPN, we extended the window from 300-800 ms to 300-1500 ms. As a result, the number of spatio-temporal features for training the classifiers using the feature extraction method in Section 3.2.2 were 72 features (12 windows $\times$ 6 channel clusters).

4.2.1 Classification Problem

Classification analysis was conducted separately on each experiment (Exp 1, Exp 2, Exp 3-location, Exp 3-color). Note that the data recorded in Experiment 3 were divided into two sets by source conditions in order to reveal any possible differences between the location and color conditions that may correspond to the ERP differences observed in Mollison and Curran (2012).

The behavioral conditions of correct item retrieval (SC and SI) and correct item rejection

(CR) were used for training classifiers. Three different two-class binary classifiers (SC-CR, SI-CR, and SC-SI) with real-valued outputs were trained to discriminate between pairs of behavioral conditions.

- SC-CR classifier

  The SC-CR classifier was expected to find a projection which maximizes the difference in the amount of information retrieved from the study phase.

- SI-CR classifier

  This classifier was designed to discriminate between correctly retrieved old items with incorrect source judgment and the correctly rejected new items.

- SC-SI classifier

  The SC-SI classifier was designed to extract the information about correctness of source memory retrieval for correctly remembered item judgments.

4.2.2 Subject-Independent Classification

A binary classifier using linear discriminant analysis (LDA) with automatic shrinkage regularization (Ledoit & Wolf, 2004) was trained to classify the feature vectors with 72 spatio-temporal features. To investigate the universal ERPs across subjects and to avoid any overfitting to the training data, the projections of the training conditions were computed using leave-one-subject-out (LOSO) cross-validation. To train with balanced classes, trials from the majority class for each subject were first randomly discarded from training to have equal numbers of trials in each class for each subject in the training data. These trials from each non-test subject were combined as the LOSO training data.

After training each classifier for each classification problem, all the trials in all conditions in the test subject were projected onto a 1-dimensional vector that was perpendicular to the classification hyperplane. These projected outputs were then transformed to probability estimates

by modeling the two classes as 1-Dimensional Gaussians (Equation 4.1). The probability scores are then computed as the estimated probabilities of belonging to class 1 (Equation 4.2).

$$N_i \sim \mathcal{N}(\mu_i, \sigma_i^2) \tag{4.1}$$

$$score = \frac{P[v \in N_1]}{P[v \in N_1] + P[v \in N_0]} \tag{4.2}$$

where μ_i and σ_i^2 are the mean and covariance of projected training data in class i=1 or 0 respectively.

4.2.3 Statistical Methods

It is advantageous to visualize EEG features utilized by the classifiers for interpreting any effects identified from the multivariate analysis and look at the consistency across training data (Haufe et al., 2014). We decided to look at the consistency of the data between subjects instead of the consistency between LOSO classifiers as the classifiers share a large portion of the data (each uses data from all other subjects). We examined the consistency between subjects of the mean difference between the two classes as preformed previously in Section 3.2.3.

4.3 Results

4.3.1 Classifier Performance

Figure 4.1 gives the ROC (receiver operating characteristic) curves for choosing different thresholds (between 0 and 1) for classification between the two selected classes for all 3 classification problems. Table 4.1 shows the area under these ROC curves and also compares them to the results of training individual classifiers for each subject using only their own data with LOTO cross-validation. The performance of LOSO classifiers showed similar results to those obtained by training classifiers for each subject individually in Noh et al. (2018). The larger discrepancy in Exp 1 SC-CR is likely due to the higher average number of trials per subject for that condition

leading to even better LOTO results. The SC-CR classifier had the best performance of the three types of classifiers. The SI-CR classifier worked better in the experiments with color source information (Exp 2 and Exp 3-col), while the SC-SI classifier had better performance in the experiments having location as source information (Exp 1 and Exp 3-loc). We also tested LDA classifiers without regularization by shrinkage which yielded similar performance to the ones with automated shrinkage.

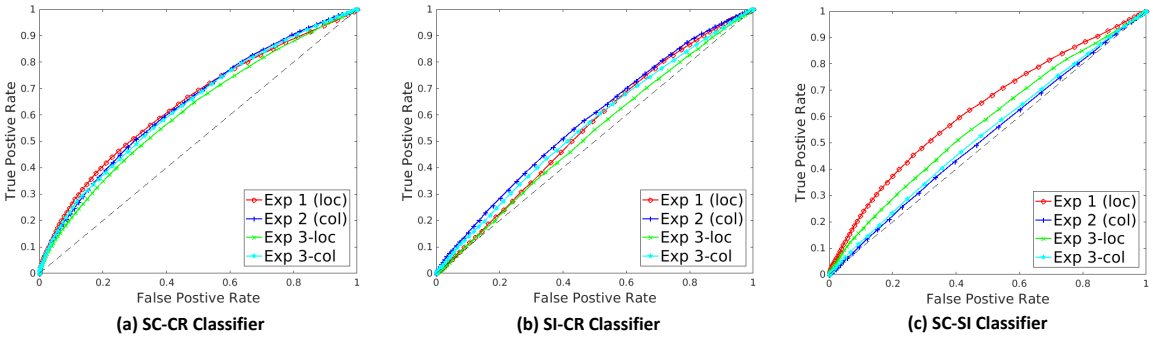


Figure 4.1. The ROC curves for the four individual datasets are given in the three different classification problems (a) SC-CR, (b) SI-CR, and (c) SC-SI.

Table 4.1. Areas under the ROC curves are given separately for different experiments and classification problems using different training paradigms. The LOTO methods trained separate classifiers for each subject using only the subject's own data. The LOTO in Noh et al. (2018) used 300-800 ms and the others used 300-1500 ms as the temporal interval.

Classifiers		LOSO	LOTO	LOTO in Noh et al. (2018)
SC-CR	Exp 1	0.6436	0.7034	0.6555
	Exp 2	0.6409	0.6727	0.6160
	Exp 3-loc	0.6141	0.6386	0.5916
	Exp 3-col	0.6355	0.6396	0.5726
SI-CR	Exp 1	0.5523	0.5377	0.5586
	Exp 2	0.5798	0.5896	0.5779
	Exp 3-loc	0.5272	0.5470	0.5264
	Exp 3-col	0.5587	0.5546	0.5376
SC-SI	Exp 1	0.6250	0.5970	0.5434
	Exp 2	0.5200	0.5151	0.5357
	Exp 3-loc	0.5707	0.5632	0.5375
	Exp 3-col	0.5322	0.5260	0.5108

4.3.2 Analysis of Classifier Scores

As discussed, each trial of EEG data can be projected onto a discriminative vector and transformed to a probability estimate of belonging to class 1 as its classifier score. The average classifier scores for different behavioral conditions show how the classifiers separate the different behaviors.

The correct item memory conditions (SC, SI, and CR) showed similar patterns across experiments (1,2,3-loc,3-col) in SC-CR classifiers where the SC trials gave the highest scores and the CR trials showed the lowest scores in Figure 4.2. Noticeable in Figure 4.2 (b) SI-CR classifiers, the SI trials were more separable from the CR trials in the color source conditions than in the location source conditions. Conversely, in Figure 4.2 (c) SC-SI classifiers, the SC trials were more different from the SI trials in the location source conditions than in the color source conditions. The patterns of these two classifiers were in accordance with the performance of the SI-CR and the SC-SI classifiers shown in Table 4.1 and Noh et al. (2018).

4.3.3 Activation Patterns

In order to investigate the consistency between subjects, the mean differences between the two classes for each subject for the three different classification problems in each experiment were calculated. The mean differences for each subject were L2-normalized, and the average of normalized values across subjects is shown in Figure 4.3. The significant clusters with $p < .05$ are shown in in Figure 4.4, and the p-values of the most significant clusters are shown in Table 4.2. The p-values in the mean differences are in accordance with the performance of the classifiers (area under ROC curves).

4.4 Discussion

In our spatio-temporal feature selection, we specifically extended (relative to Noh et al. (2018)) the temporal window to 1500 ms in order to capture the possible late posterior negativity

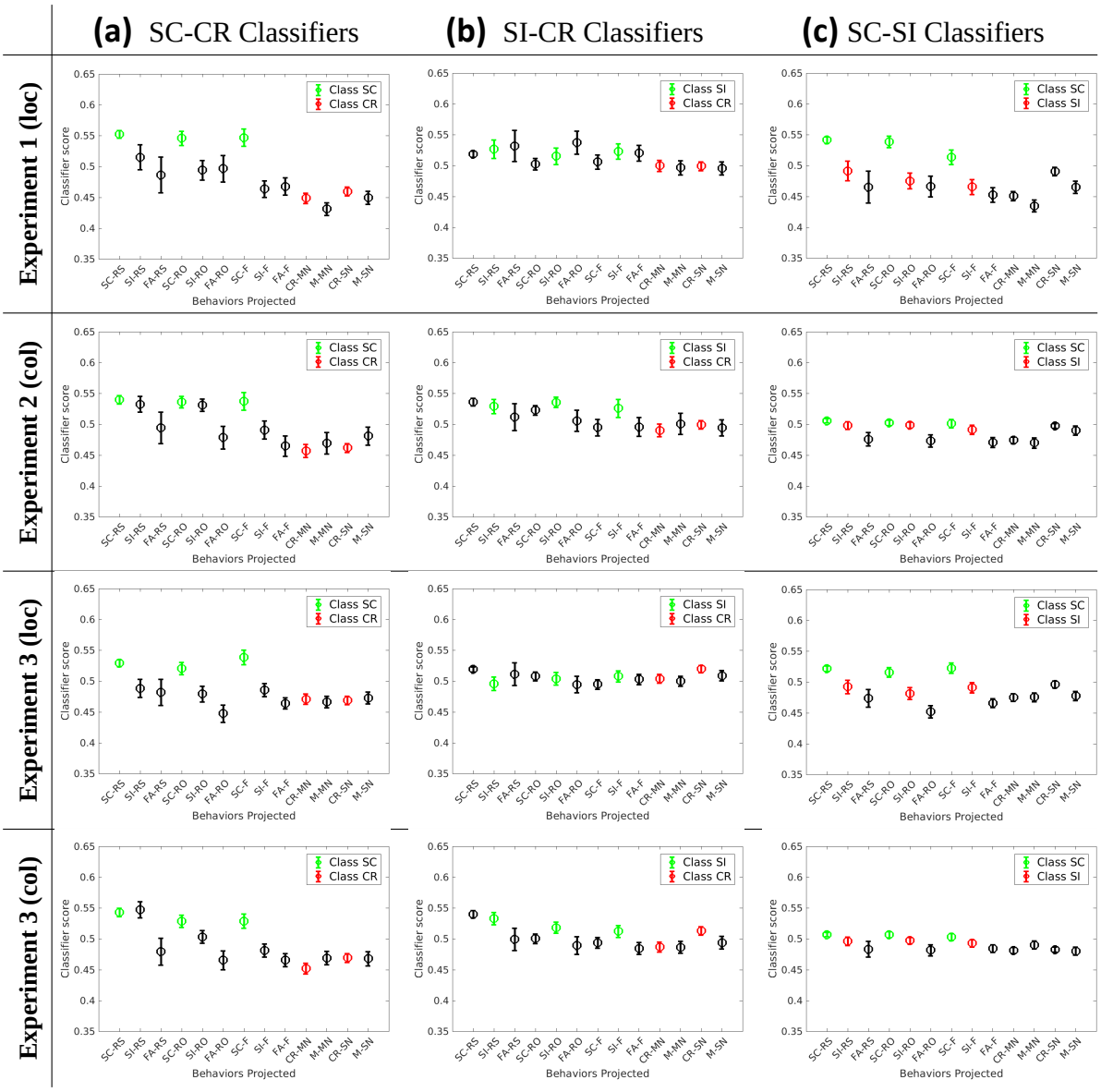


Figure 4.2. The scores of projected trials in different behaviors using projection functions from (a) SC-CR classifiers, (b) SI-CR classifiers, and (c) SC-SI classifiers are given separately for four individual datasets.

(LPN) in the recorded data. In Table 4.1, the three different classifiers trained on LOTO basis using features from 300 to 1500 ms outperformed the ones using features from 300 to 800 ms suggesting that the features between 800 ms and 1500 ms are informative for our memory classification problems. The subjects were not allowed to respond until 1500 ms after stimulus

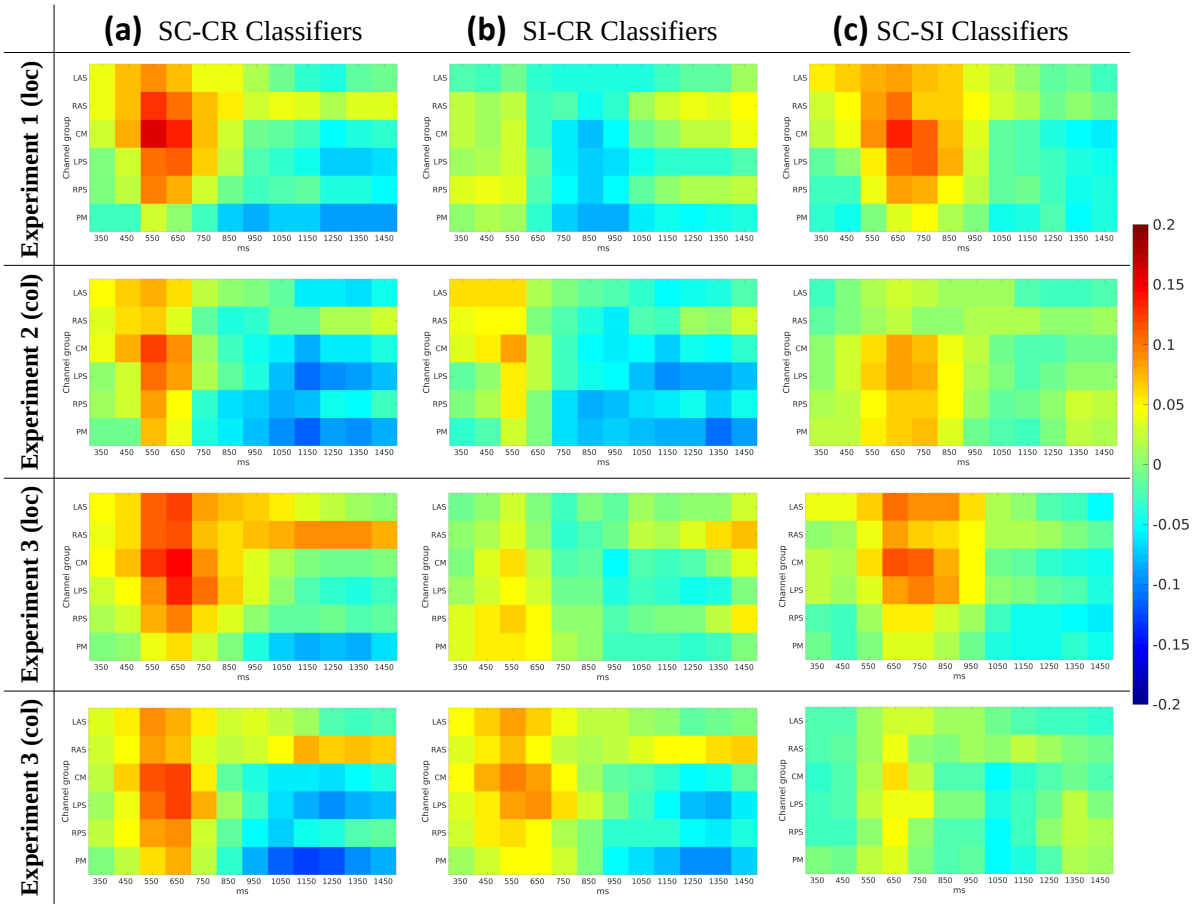


Figure 4.3. The patterns are the average of normalized mean difference between two classes for each subject.

Table 4.2. P-values for the most significant cluster in Figures 4.3/4.4

Indiv Class Diff	SC-CR	SI-CR	SC-SI
Exp 1	0.0005	0.0081	0.0007
Exp 2	<0.0001	0.0002	0.0092
Exp 3-loc	<0.0001	0.0750	0.0004
Exp 3-col	0.0002	0.0011	0.1738

presentation, and response assignments for the keys were counterbalanced across participants, so response related effects in LOSO are minimized. In Figure 4.4 (a), the late posterior effect was consistent in the SC-CR classification problem in all experiments except for Exp 3-loc (where the tendency is visible in Figure 4.3 but does not rise to significance using our cluster test). In the SC-SI classification problem in Figure 4.4 (c), the consistent wide-spread patterns in the location

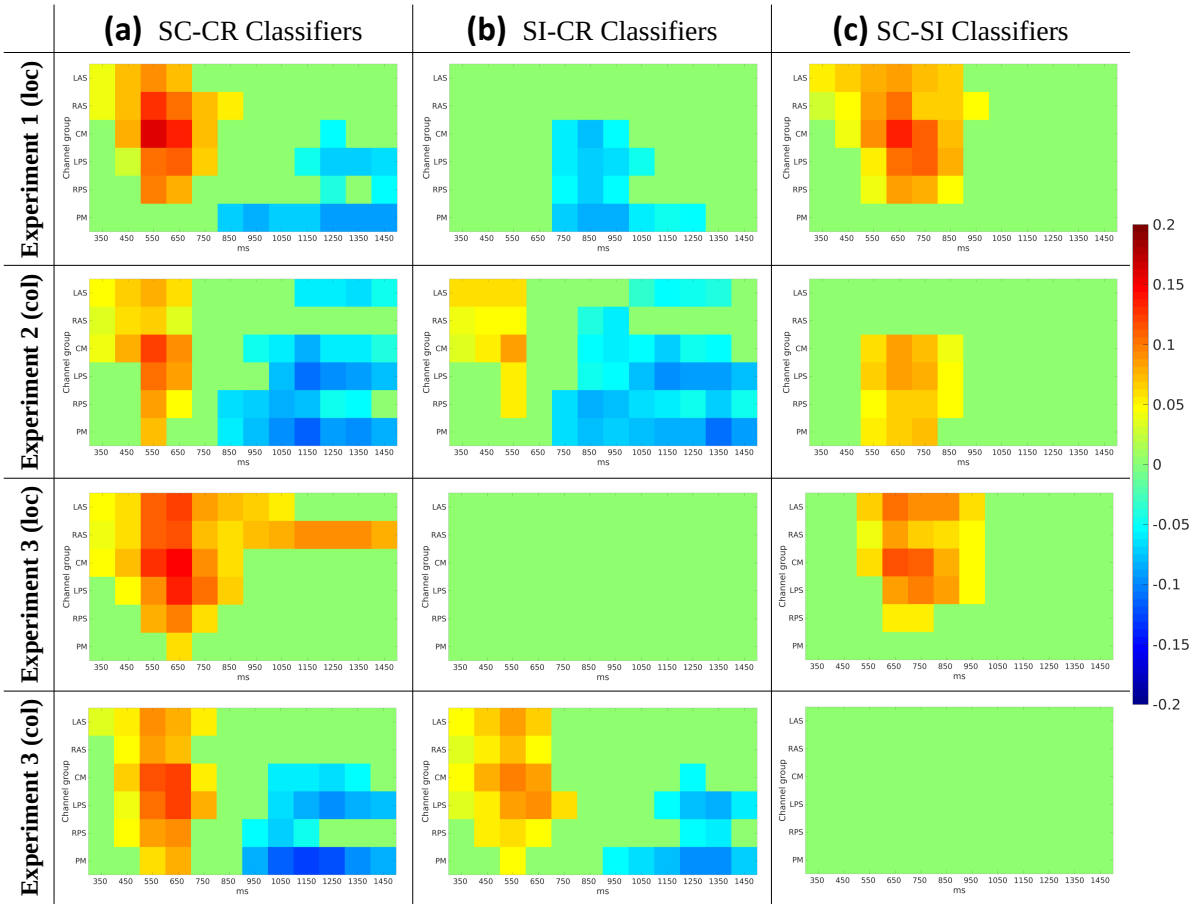


Figure 4.4. The significant clusters of features ($p < .05$) in the patterns of the average of normalized mean difference between two classes for each subject.

source conditions after 800 ms could explain why the extension of the temporal window leads to better performance in SC-SI classifiers in Experiments 1 and 3-loc.

In this chapter, we showed that it is possible to predict memory retrieval based on single-trial EEG on new subjects not in the training data. The LOSO classifiers had similar performance to LOTO classifiers in Table 4.1 trained on individual subjects. Except for Exp 1 SC-CR, the results were within a few percent of the analogous LOTO results. For the SC-SI and SI-CR classifications the LOSO classifications were often slightly better than LOTO results while for the SC-CR classifier the individualized classifiers were better. The successful prediction of memory retrieval by the LOSO method implies that single-trial EEG classification could be applied to subjects without recording their EEG data and training personalized classifiers.

Table 4.3. Number of trials in each class used by 3 classifiers in Experiment 1. Dash represents the classifier wasn't trained due to number of trials less than 25 in either class.

LOSO/Sub	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
SC-CR	2447	2517	2417	2551	2445	2438	2459	2497	2530	2479	2467	2472	2462	2438	2412	2478	2471	2456	2515	2475	2515	2480	2441	2470	2429	2414
SI-CR	1121	1153	1118	1154	1093	1114	1148	1135	1153	1132	1095	1097	1119	1129	1135	1086	1126	1106	1123	1120	1144	1116	1123	1134	1130	1121
SC-SI	1121	1153	1118	1154	1093	1114	1148	1135	1153	1132	1095	1097	1119	1129	1135	1086	1126	1106	1123	1120	1144	1116	1123	1134	1130	1121
LOTO/Sub	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
SC-CR	120	50	150	-	122	129	108	70	37	88	100	95	105	129	155	89	96	111	52	92	52	87	126	97	138	153
SI-CR	48	-	51	-	76	55	-	34	-	37	74	72	50	40	34	83	43	63	46	49	25	53	46	35	39	48
SC-SI	48	-	51	-	76	55	-	34	-	37	74	72	50	40	34	83	43	63	46	49	25	53	46	35	39	48

Table 4.4. Number of trials in each class used by 3 classifiers in Experiment 2. Dash represents the classifier wasn't trained due to number of trials less than 25 in either class.

LOSO/Sub	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
SC-CR	2387	2399	2399	2360	2340	2384	2383	2308	2386	2272	2285	2396	2334	2332	2296	2373	2332	2332	2375	2324	2377	2327	2312	2372	2288	2395	2342	2308
SI-CR	1815	1827	1827	1836	1791	1814	1834	1792	1814	1785	1776	1824	1762	1782	1733	1812	1778	1822	1803	1771	1810	1793	1759	1800	1768	1823	1771	1752
SC-SI	1948	1994	1988	2004	1959	1982	2002	1960	1973	1953	1944	1965	1924	1950	1901	1980	1946	1990	1914	1939	1978	1961	1927	1966	1936	1967	1939	1920
LOTO/Sub	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28
SC-CR	47	35	35	74	94	50	51	126	48	162	149	38	100	102	138	61	102	102	59	110	57	107	122	62	146	39	92	126
SI-CR	47	35	35	26	71	48	28	70	48	77	86	38	100	80	129	50	84	40	59	91	52	69	103	62	94	39	91	110
SC-SI	82	36	42	26	71	48	28	70	57	77	86	65	106	80	129	50	84	40	116	91	52	69	103	64	94	63	91	110

In our previous work using LOTO training for each subject, only the classification problems with enough trials ($\geq$25 in each class) could be investigated due to the limited numbers of trials of certain behaviors for some subjects. Sufficient trials are necessary to fit the covariance matrices used in linear discriminant analysis. Although the issue could be mitigated somewhat by using a higher shrinkage parameter during training, over regularization will also lead to decreased accuracy. In contrast, the trials of a behavior in the LOSO classification were concatenated across training subjects. (Tables 4.3 through 4.6 show the number of trials for each experiment and classification task for our LOSO and LOTO classifiers). Therefore, the training data has many more trials in each class compared to LOTO training. LOSO training provides an opportunity to investigate the relationships between behaviors with few trials and consistent features across subjects.

Table 4.5. Number of trials in each class used by 3 classifiers in Experiment 3-loc. Dash represents the classifier wasn't trained due to number of trials less than 25 in either class.

LOSO/Sub	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
SC-CR	2253	2186	2218	2224	2235	2300	2279	2274	2288	2261	2267	2201	2254	2238	2204	2224	2249	2294	2279	2170	2205	2311	2214	2284	2280	2233
SI-CR	1150	1163	1173	1106	1135	1159	1138	1133	1147	1153	1131	1125	1179	1159	1134	1172	1173	1153	1177	1165	1099	1170	1164	1143	1139	1160
SC-SI	1359	1372	1382	1315	1344	1364	1308	1331	1348	1362	1340	1334	1388	1368	1343	1381	1382	1347	1386	1374	1308	1330	1373	1297	1320	1369
LOTO/Sub	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
SC-CR	84	151	119	113	102	37	58	63	49	76	70	136	83	99	133	113	88	43	58	167	132	26	123	53	57	104
SI-CR	46	33	-	90	61	37	58	63	49	43	65	71	-	37	62	-	-	43	-	31	97	26	32	53	57	36
SC-SI	46	33	-	90	61	41	97	74	57	43	65	71	-	37	62	-	-	58	-	31	97	75	32	108	85	36

Table 4.6. Number of trials in each class used by 3 classifiers in Experiment 3-col. Dash represents the classifier wasn't trained due to number of trials less than 25 in either class.

LOSO/Sub	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
SC-CR	2339	2263	2287	2292	2283	2358	2353	2329	2358	2291	2325	2261	2312	2327	2251	2271	2307	2350	2313	2247	2262	2348	2306	2377	2342	2298
SI-CR	1748	1717	1759	1703	1724	1774	1762	1738	1767	1724	1734	1692	1756	1761	1691	1752	1755	1759	1766	1723	1679	1757	1754	1786	1751	1743
SC-SI	2062	2043	2085	2029	2050	2100	2023	2061	2049	2050	2034	2018	2082	2087	2017	2078	2081	2062	2092	2049	2005	2048	2080	2024	2047	2069
LOTO/Sub	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
SC-CR	63	139	115	110	119	44	49	73	44	111	77	141	90	75	151	131	95	52	89	155	140	54	96	25	60	104
SI-CR	63	94	52	108	87	37	49	73	44	87	77	119	55	50	120	59	56	52	45	88	132	54	57	25	60	68
SC-SI	75	94	52	108	87	37	114	76	88	87	103	119	55	50	120	59	56	75	45	88	132	89	57	113	90	68

In order to reveal the latent relationships between the 3 selected classification problems, we investigated how well classifiers trained on each classification problem were able to solve all 3 classification problems. For instance, we have already examined how well the SC-CR classifier is able to separate the SC vs CR trials, but we can also see if it can separate the SC from the SI trials and the SI from the CR trials. Likewise we can ask similar questions using the SI-CR and SC-SI classifiers. The areas under the ROC curves were calculated for the scores from the projections of each pair of classes onto vectors perpendicular to each classification problem as shown in Table 4.7. In the table, the different projection functions/directions appear as different columns and the classification problems (data) appear as rows. The first four rows (SC-CR) show that the SC vs CR trials are best separated by the SC-CR classifiers but are somewhat separable by the SI-CR and SC-SI classifiers/projections. In particular the experiments with spatial source (Exp1 and Exp3-loc) have their SC and CR trials well separated by the SC-SI classifiers. The SI vs CR trials were actually better separated by the SC-CR classifiers (except for Exp1 which is close). The SC vs SI trials were fairly similarly separated by the SC-CR classifiers and the SC-SI classifiers. These findings are in accordance with the ERP differences observed in Mollison and Curran (2012) and the distribution of scores for each behavior when projected onto the different discriminant directions for each classifier in Figure 4.2. We conclude that an SC-CR classifier trained on other subjects is able to well separate SC, SI, and CR trials in another subject.

Table 4.7. Areas under ROC curves calculated based on the scores computed from projections of behaviors with different classifiers

Behaviors \ Classifiers		SC-CR	SI-CR	SC-SI
SC vs CR	Exp1	0.6436	0.5228	0.6065
	Exp2	0.6409	0.5722	0.5498
	Exp 3-loc	0.6141	0.5311	0.5826
	Exp 3-col	0.6355	0.5523	0.5765
SI vs CR	Exp1	0.5446	0.5523	0.4828
	Exp2	0.6174	0.5798	0.5303
	Exp 3-loc	0.5465	0.5272	0.5122
	Exp 3-col	0.5964	0.5587	0.5451
SC vs SI	Exp1	0.5987	0.4713	0.6250
	Exp2	0.5260	0.4927	0.5200
	Exp 3-loc	0.5700	0.5042	0.5707
	Exp 3-col	0.5410	0.4940	0.5322

4.5 Chapter Acknowledgements

Chapter 4, in full, is a reorganized version of the following publication in IEEE International Conference on Bioinformatics and Biomedicine, 2018. Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. "Single-trial EEG predicts memory retrieval using leave-one-subject-out classification", 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2613-2620, 2018. The dissertation author was the primary author of this publication.

Chapter 5

Control for Confidence Reveals Familiarity

5.1 Introduction

Electroencephalography (EEG) has been widely used to identify neural substrates and cognitive processes in recognition memory studies for its noninvasive temporal sensitivity. The event-related potential (ERP) method that takes time-locked averages of multiple trials in EEG data is most commonly used. The frontal old/new effect (also called the FN400) is a negative-going ERP observed in the frontal electrodes that peaks around 400 ms post stimulus. The FN400 goes more negative for less familiar items (Curran, 2000; Curran & Hancock, 2007) but disassociates from the amount of recollected episodic information (Curran & Cleary, 2003; Rugg & Curran, 2007). Hence, the FN400 is considered as a familiarity-related ERP. The parietal old/new effect, also called the late positive component (LPC), is another ERP that is positive-going and peaks over the parietal scalp between 500 and 800 ms. The LPC shows greater amplitude for correctly identified old items (hits) as opposed to new items (correct rejections) (Rugg et al., 1998; Curran, 2000; Wilding, 2000) and positively correlates with the amount of information retrieved from the study episode (Wilding & Rugg, 1996; Vilberg, Moosavi, & Rugg, 2006). Therefore, the LPC is thought to reflect recollection. Another memory-related ERP is the late posterior negativity (LPN). The LPN emerges at approximately 800 ms post stimulus and goes more negative for correct old than new responses, irrespective of the accuracy of the

retrieved information (Johansson & Mecklinger, 2003; Friedman et al., 2005; Herron, 2007).

In the remember-know (RK) paradigm, it is difficult to separate the effects of memory from those of any decision confidence component because the difference between remember and know judgements could be derived from both effects simultaneously (Tulving, 1985; W. Donaldson, 1996; Yonelinas et al., 2002). Likewise the difference between responses to Know and New items may reflect both differences in familiarity and confidence. ERP studies have revealed differences between high and low confidence in both old and new memory judgements around 600-800 ms over parietal scalp (Addante et al., 2012; Wynn et al., 2019), but decision confidence as a similar process for both old and new items has rarely been studied in ERP studies. However, in a single-neuron recording study of posterior parietal cortex (Rutishauser et al., 2018), confidence-selective cells encoding retrieval confidence for both old and new stimuli were identified.

Recently, multivariate pattern classification (MVPC) methods applied to EEG data recorded in episodic memory tasks have helped elucidate brain activity during encoding (Noh et al., 2014; Anderson et al., 2016) and decoding (Noh et al., 2018; Liao et al., 2018). Our goal was to separate, as much as possible, the familiarity/memory based component from any confidence based component in familiarity judgements.

5.2 Methods

In this chapter, the datasets, experimental paradigms for data collection, and the data preprocessing remained the same like those in Chapter 2. However, we only focused on Experiment 1 and Experiment 2 to avoid the non-stationarity of the data between the two sessions (due to electrode position changes, impedance changes, or changes in brain-state) in Experiment 3.

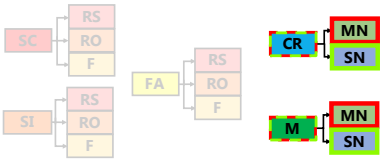


Figure 5.1. Green and red box outlines indicate the positive and negative behaviors for training the SN vs. MN classifier.

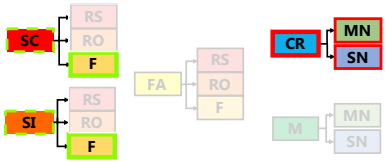


Figure 5.2. Green and red box outlines indicate the positive and negative behaviors for training the F vs. CR classifier.

5.2.0.1 Classification

Classification analysis was conducted separately for Exp 1 and Exp 2 in order to reveal any possible difference between the location source and color source experiment responses that may correspond to the differences in ERPs observed by Mollison and Curran (2012).

Two binary classifiers were trained to discriminate between pairs of behavioral conditions:

- SN vs. MN classifier

  The SN and MN class included both new responses (see Figure 5.1). This classifier was designed to distinguish different levels of confidence when we excluded the time window related to the familiarity-related ERP FN400.

- F vs. CR classifier

  The F class included SC-F and SI-F, and the CR class consisted of CR-SN and CR-MN (see Figure 5.2). This classifier was trained to identify the familiarity process in the memory retrieval task.

5.2.1 Training Classifiers

Spatio-temporal features of the ERPs were extracted using the feature extraction method in Section 3.2.2. The period of 600 to 1500 ms after probe item presentation in the recognition

phase was considered for the SN vs. MN classifier in order to avoid the familiarity effect (FN400). For the F vs. CR classifier, the post-item interval of 300 to 1500 ms was considered in order to cover all discussed memory-related ERPs. By averaging over 100 ms non-overlapping windows, overall spatio-temporal features extracted were 54- and 72-dimensional feature vectors for each trial for the SN vs. MN and F vs. CR classifier, respectively.

The type of classifier trained to classify the feature vectors was the linear discriminant analysis (LDA) classifier with automatic shrinkage regularization as presented by Schäfer and Strimmer (2005) based on the approach of Ledoit and Wolf (2004). Leave-one-subject-out (LOSO) cross-validation (Liao et al., 2018) was utilized for training the classifier to avoid over-fitting and exploit the consistent spatio-temporal features across subjects. The trials from the non-test subjects were combined as the LOSO training data. The data were centered for each class and merged to obtain a more reliable shared covariance matrix for determining the classifier. A linear classifier learns a hyperplane to best separate the two classes. We refer to the vector perpendicular to the separating hyperplane and pointing in the direction of the first-named (or positive trained) class as the discriminant vector. After training, all the data from the test subject, including conditions not in the two training classes, were projected (using the dot product) onto the discriminant vector to determine a signed distance, or projection, from the classification hyperplane.

5.2.2 Differentiate Conditions not Trained

After training the classifiers, the performance of each classifier can be evaluated using metrics computed on the projection of trials of the trained conditions from the test subjects. Similarly, trials from untrained classes can be projected and compared (Noh & de Sa, 2014). In this chapter, we show that in addition, how closely aligned the difference between any two selected conditions is with the classifier determined by the training conditions could also be assessed using metrics on the projections of the selected conditions onto the discriminant vector of the classifier. We used area under the ROC curve (AUROC) as our metric to avoid bias due

to imbalance of conditions. For example, if for paired conditions X and Y, the AUROC of the classification of their projections onto the SN vs. MN classifier are significantly above 0.5, this would indicate that X and Y are somewhat aligned with the classification boundary for CR-SN and CR-MN, with X more like CR-SN, and Y more like CR-MN.

5.2.3 Condition-Controlled Classifier Training

The AUROC of F vs. CR projections on the SN vs. MN classifier was significantly below 0.5 as shown in the Results section, which meant the F vs. CR classifier trained could have been trained based on both the confidence difference and the familiarity process (see Discussion in Section 5.4). In order to control the confidence level of F and CR and because there were more CR trials than F trials in general, we sorted the CR trials in the training data based on their projections onto the trained SN vs. MN confidence classifier, selected the CR trials with the smallest values, and accumulated CR trials in a bottom-up manner until their average was the same as the average of the projections of the F trials in the training data. Then the F and selected CR trials were used as the new training data for training a F vs. CR classifier with decision confidence (as defined by projections from the SN vs MN classifier) controlled.

5.2.4 Visualization of Consistent EEG Features

We also examined the consistent (across subjects) and important EEG features for both classification problems by calculating the mean difference between the two classes for each subject as performed in Section 4.2.3. The only difference was the time window here was 250-1450 ms.

5.3 Results

5.3.1 Classifier Performance

For this analysis, only subjects having at least 10 trials in both test classes were included. Table 5.1 shows the numbers of subjects in the analysis and the AUROC of the SN vs. MN and

Table 5.1. AUROCs and accuracies calculated based on the scores computed from projections of behaviors from different classifiers. RS and ConfMatched refer to SC-RS and confidence matched, respectively. The number of subjects with at least 10 trials in both test classes and the number of total subjects are given as the numerator and the denominator, respectively.

Behaviors \ Classifiers		SN vs. MN		F vs. CR		F vs. CR (ConfMatched)	
		AUROC	Acc.	AUROC	Acc.	AUROC	Acc.
SN vs. MN	Exp 1 (24/26)	0.5564**	0.5421**	0.4456**	0.4575**	0.4796	0.4884
	Exp 2 (26/28)	0.5997**	0.5653**	0.4146**	0.4537**	0.4528**	0.4639**
F vs. CR	Exp 1 (25/26)	0.4434*	0.4637*	0.5793**	0.5465**	0.5615**	0.5429**
	Exp 2 (24/28)	0.4400*	0.4576*	0.5782**	0.5405*	0.5585**	0.5455**
RS vs. F	Exp 1 (25/26)	0.6552**	0.6126**	0.4298**	0.4606**	0.5045	0.5002
	Exp 2 (24/28)	0.6277**	0.5829**	0.4797	0.4909	0.5305	0.5027

** $p < 0.01$,* $p < 0.05$

F vs. CR classifiers in Experiments 1 and 2. The accuracy of the classifiers were calculated on balanced test data. We calculated the average AUROC and accuracy for each subject. The non-parametric two-sided Wilcoxon Signed Rank Test was applied to the subject AUROC and accuracy data to determine significance relative to chance performance. While the SN vs. MN classifier performs significantly above chance, it does better in Exp 2 than in Exp 1. This reflects the difference between the two source conditions and indicates the confidence gap could be larger for new responses for the color condition. Performance of the F vs. CR classifiers are significantly above chance and approximately the same for both Exp 1 and 2.

5.3.2 Projections from the Classifiers

Figure 5.3 (a) and (b) show the projections of all the behaviors from the SN vs. MN classifier in Experiment 1 and 2, respectively. In Exp 1, SC-RS receives the highest value when projected on the SN vs. MN classifier, implying that this classifier classifies largely on decision confidence, with high decision confidence common to SC-RS and SN. (It also appears that any possible negative memory or familiarity component that might be more present in MN than SN is minimal). In Exp 2, SC-RS and SC-RO are also projected to high values on the SN vs. MN classifier while the projections of SC-F and SI-F are both low in both experiments. The same relationship between SC-RS and SC&SI-F could also be found in the last two rows in Table 5.1.

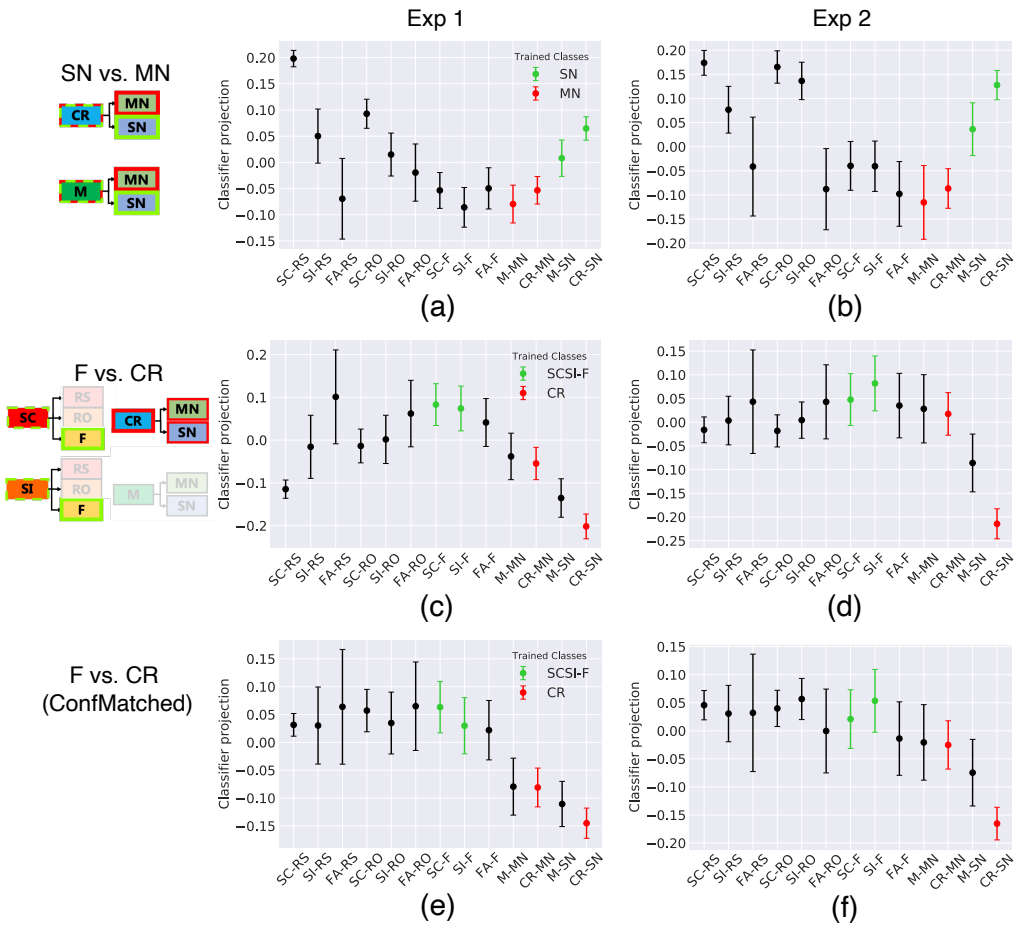


Figure 5.3. The average projection and the 95% CI of behaviors from the trained classifiers in Exp 1 and 2. The positive and negative trained classes are plotted in green and red, respectively. The classes shown in black were not trained.

In Figure 5.3 (c) and (d), the F vs. CR classifier projects SC-RS to a lower value that of SC-F and SI-F, especially in Exp 1. Table 5.1 shows the significant difference between the projected SC-RS and the projected SC&SI-F from the F vs. CR classifier in Exp 1.

5.3.3 Activation Patterns

The spatio-temporal features that are consistent across subjects in each classification problem for Exp 1 and 2 calculated using the cluster-based multiple comparison method are shown in Figure 5.4. Note that the significance test for SN vs. MN was only performed for the features after 600 ms to match the time frame used for training the SN vs. MN classifier. The

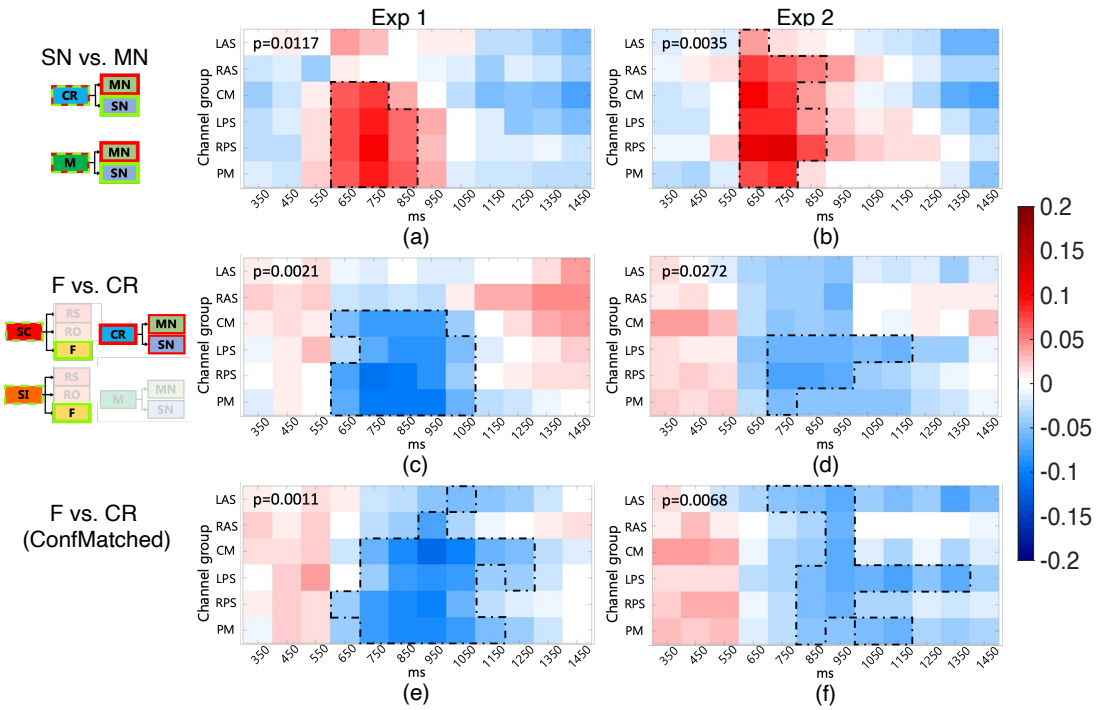


Figure 5.4. Mean difference between training classes of each classification problem in Exp 1 and 2. The areas surrounded by the dotted lines are the top positive/negative clusters with p-value $< .05$. The p-value at the top-left corner of each panel is calculated based on the cluster with the largest absolute t-stat.

peak of the positive clusters in SN vs. MN (in Figure 5.4 (a) and (b)) is around 700-800 ms in both experiments. Note that the positive cluster in SN vs. MN is highly overlapped with the significant cluster in F vs. CR in Exp 1 (Figure 5.4 (c)). The wide overlap and opposite sign also reflect the AUROCs that are significantly below 0.5 when using F vs. CR and SN vs. MN to classify each other in Exp 1 in Table 5.1.

5.4 Discussion

In this chapter, we first trained an SN vs. MN classifier to distinguish SN from MN based on their decision confidence level difference and an F vs. CR classifier hoping to reveal the familiarity process. Observing the projections of all behaviors from the SN vs. MN classifier and from the F vs. CR classifier, we found that the F vs. CR classifier reflected classification based not only on familiarity but also confidence. Hence, we further trained an F vs. CR classifier with

confidence control.

5.4.1 Difference between SN and MN

There are two potential factors that the SN vs. MN classifier utilized to differentiate SN and MN: one is the confidence difference, and the other is familiarity or memory strength as in the dual-process models (Wixted, 2007; Wixted & Mickes, 2010). If it was simply familiarity or memory strength that the SN vs. MN classifier classified on, F and RS should have been projected lower (beyond MN). Nevertheless, what is shown in Table 5.1 is that the AUROCs of RS vs. F are significantly above 0.5 on the SN vs. MN classifier in both location and color conditions. Moreover, SN vs. MN was deliberately trained on features excluding those in the time period of the FN400. These support the idea that SN vs. MN is primarily classifying based on confidence rather than memory, and is consistent with Remember responses having high confidence, and Know responses lower confidence (Tulving, 1985).

5.4.2 Confidence Component Revealed in F vs. CR

In Figure 5.3 (c) and (d), SN receives lower projected values than MN on the F vs. CR classifier in both Exps 1 and 2. This is reasonable if the classification was performed based on familiarity. However, in Figure 5.3 (c), the projection of SC-RS is the lowest among all the old judgements. In fact, the shape of the projections of the behaviors in Figure 5.3 (c) is like the vertically mirrored version of the shape in Figure 5.3 (a). Moreover, the peak of the negative cluster in the activation pattern of F vs. CR (Figure 5.4 (c)) overlaps with the peak of the positive cluster in the pattern of SN vs. MN (Figure 5.4 (a)) at around 700 to 800 ms for the location condition. Based on these observations and because F is generally associated with low confidence, the F vs. CR classifier in the location condition could have incorporated some (negative) confidence component and thus give higher scores to the responses with less confidence (more like F). To overcome the confidence effect in the familiarity classifier, we controlled the confidence of the F and CR classes and trained a new confidence-matched F vs.

CR classifier (see Methods: Condition-Controlled Classifier Training).

5.4.3 Confidence Matched F vs. CR

The average of the projections of MN responses from the F vs. CR classifier becomes negative when confidence is controlled for, which better matches the desired output of a F vs. CR classifier. Also, with confidence matching, the new F vs. CR classifier for both conditions now projects RS responses above 0 and at the same level as F responses as shown in Figure 5.3. The RS trials do not have higher projections than the F trials on the confidence matched F vs. CR classifier either because the recollection and familiarity trials have similar familiarity once confidence is controlled for, or possibly due to some remnant of the confidence effect. Future experiments could further investigate and compare the "familiarity" strength of familiar and recollected items.

The activation patterns of the original F vs. CR without confidence match show significant negative clusters which peak around 700 to 800 ms for both location and color conditions in Figure 5.4 (c) and (d). With confidence matching of F and CR, the peaks are now later (around 900 to 1000 ms) and farther from the peaks in the SN vs. MN classifier. This late negative component also appears to consistently (across subjects) extend more frontally, once we match for confidence. Exploring this finding would be an interesting area for future careful experimental design.

The same confidence control method for F vs. CR classifier was also tested with using the CR-SN vs. CR-MN confidence classifier defined in the following chapter, and the results of RS and F having similar familiarity strength remain unchanged.

5.5 Chapter Acknowledgements

Chapter 5, in full, is a reorganized version of the following publication in the Proceedings of the Annual Meeting of the Cognitive Science Society, 2021. Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. "EEG reveals familiarity by controlling confidence

in memory retrieval", Proceeding of the Annual Meeting of the Cognitive Science Society, 43, 2021. The dissertation author was the primary author of this publication.

Chapter 6

Remember-Know Responses: Difference in Confidence, Source Memory, and Item Memory

6.1 Introduction

6.1.1 ERPs in EEG Data Analysis

Electroencephalography (EEG) is the most commonly used neuroimaging modality for studying neural substrates and cognitive processes because of its temporal sensitivity and non-invasiveness. Event-related potentials (ERP), the time-locked average EEG of a number of trials, are the most common measure of the brain response in EEG to different stimuli or events. By showing the significant difference between the ERPs of conditions in either time windows or locations on the scalp, researchers could identify the EEG variations elicited by certain cognitive processes and even achieve source localization.

In recognition memory EEG studies, several types of ERPs have been found associated with different memory processes. The frontal old/new effect (called FN400) is a frontally distributed and negative-going ERP peaking around 400 ms post stimulus onset. The FN400 shows greater negative amplitude for correct rejections than hits; in other words, it goes more negative for less familiar items. Meanwhile, the FN400 typically does not vary for different amounts of recollected episodic information (Curran, 2000; Mecklinger, 2006; Rugg & Curran,

2007; Tsivilis et al., 2015). Therefore, it is understood as a neural correlate of familiarity and item memory. Another type of ERP is the parietal old/new effect. It is a positive-going ERP and peaks over the parietal scalp between 500 and 800 ms. Because it emerges slightly later than the FN400 and is a positive deflection, it is also referred to as the late positive component (LPC). The LPC is typically lateralized on the left and shows greater amplitude for the correctly identified old (hits) items compared to new (correct rejections) items (Rugg et al., 1998; Curran, 2000; Wilding, 2000). Additionally, this ERP has been found to index the amount of information retrieved from the study episode (Wilding & Rugg, 1996; Curran, 2000; Vilberg et al., 2006; Woroch & Gonsalves, 2010); hence, it is thought to reflect recollection and source memory. The late posterior negativity (LPN) is another type of memory-related ERP. It emerges approximately 800 ms post stimulus, even later than the LPC. It is considered to reflect a functionally heterogeneous effect with interpretations ranging from episodic information searching to response fluency (Johansson & Mecklinger, 2003; Herron, 2007; Friedman et al., 2005; D. I. Donaldson & Rugg, 1998). The LPN yields greater negative amplitude for correct old than new responses, irrespective of the accuracy of the searched and retrieved information.

However, ERP analysis would fail to separate all the underlying cognitive processes when the differences between conditions consist of multiple processes and the ERPs corresponding to these processes overlap temporally and spatially. As a result, the observed ERP can not be correctly associated with separate underlying processes. In (?, ?), recollection was found to be a combination of item memory and source memory, and simply taking the ERP difference between the conditions of retrieving source or not would result in the ERP showing a combined difference resulting from possible differences in both item memory and source memory. Similarly, in the debate over whether the remember-know (RK) distinction (Tulving, 1985) should be interpreted with a single-process or a dual-process model (W. Donaldson, 1996; Dunn, 2004; Yonelinas et al., 2002; Wixted & Stretch, 2004), it is difficult to separate the effect of memory and confidence in the ERPs as the remember judgements are generally considered to be high confidence and high memory strength judgements, whereas the know judgements are considered lower confidence

and possibly reflecting lower memory strength.

There are several approaches to mathematically isolate the latent components in ERPs (Luck & Gaspelin, 2017). Spatial principal component analysis (PCA) could quantify the magnitude of the underlying component at each electrode sites, temporal PCA could decompose the temporally overlapped underlying components across different conditions, and spatio-temporal PCA attempts to extract the underlying components based on both temporal and spatial patterns (Parra, Spence, Gerson, & Sajda, 2005; Dien, Beal, & Berg, 2005). However PCA is an unsupervised algorithm that finds perpendicular axes that explain the most variance. This method will fail to find the functionally underlying components if these components are not perpendicular to each other. Independent component analysis (ICA) and source localization are more powerful unsupervised methods for identifying latent components (Makeig, Onton, et al., 2011). However, both methods require significant manual work to select meaningful components or sources.

6.1.2 EEG, Single-Trial Classifier, LOSO

Recently, multivariate pattern classification (MVPC) methods applied to neural activity recorded with EEG have shed light upon representations in brain activity during encoding (Jafarpour et al., 2014; Noh et al., 2014; Anderson et al., 2016; Ratcliff et al., 2016) and retrieval (Jafarpour, Horner, Fuentemilla, Penny, & Duzel, 2013; Johnson, Price, & Leiker, 2015; Liao et al., 2018; Noh et al., 2018; Kerrén, Linde-Domingo, & Hanslmayr, 2018) of episodic memory. In (Liao et al., 2018), the subject independent classifier was trained with a dimensionality reduction method (Noh & de Sa, 2014) and leave-one-subject-out (LOSO) cross-validation by combining trials of non-test subjects for training, and the performance was consistent with a leave-one-trial-out (LOTO) intra-subject classifier (Noh et al., 2018). Such methods provide an opportunity to investigate the brain activity across conditions and train a classifier even when the trial count is low for the training class for a subject (Noh & de Sa, 2014). In ERP studies and intra-subject classification, these data were usually ignored or combined with other behaviors in order to acquire a reasonable number of trials to average in the ERP for analysis and reliable

classification hyperplane(s), respectively.

6.1.3 Our Goal

Our study aimed to utilize pattern classifiers for different classification problems to separate and represent different components in memory retrieval. Different behaviors were projected onto the discriminant vector established in each classification problem. The separability of projected behaviors (based on area under the ROC curve) revealed whether the behaviors were distinguishable from one another using only the features that distinguished the original two classes separated by the classifier. Also, the projections of data onto different classifiers could be used for selecting unbiased training data for the classification problem of interest that is balanced with respect to another aspect. The EEG data were previously collected by (Mollison & Curran, 2012) from two separate visual memory task experiments with two types of source information. Spatial information (the location of the item) was considered in Experiment 1, and frame color information (the color of the external frame) was of interest in Experiment 2. Data collected from these two experiments were used to conduct multivariate analysis for features of interest via pattern classifiers. Finally, the linear regression models were applied to show that the remember and know differences can be considered as a mixture of a decision confidence difference, a source memory difference, and an item memory difference.

6.2 Selected Behaviors for Training the Classifier

As recollection (remember source responses) reflects high memory strength and high confidence, the difference in confidence can not be straightforwardly shown by comparing recollection to familiarity (familiar responses) in ERP studies because the differences could be accounted for by differences due to source memory, item memory, and confidence. In this section, four classifiers were trained to differentiate the selected behaviors in order to reveal the differences between recollection and familiarity in confidence, source memory, and item memory separately using the pattern classifiers. The confidence classifier was trained to identify the

confidence level of the response, the source memory classifier was trained to reveal the source memory strength of the response, the item memory classifier was trained to reveal the memory strength corresponding to the item of the response, and the R-K classifier was trained to reveal the difference between R and K judgments in the R-K paradigm.

6.2.1 Materials

Electroencephalography recordings for the this chapter came from three separate visual memory task experiments (Mollison & Curran, 2012) as shown in Section 2. Also, the data preprocessing were the same as those in Chapter 2. Only Experiment 1 and 2 were considered in order to avoid the nonstationarity in Experiment 3. The time window of interest was 250-1450 ms, and the number of spatio-temporal features extracted using the feature extraction method in Section 3.2.2 were 72 features (12 windows $\times$ 6 channel clusters).

6.2.2 Classification Problem

While there could be a difference in item memory, our hypothesis was that the major difference between the correct item rejection (CR) with different levels of confidence (CR-SN and CR-MN) was substantially confidence; we test the validity of this hypothesis in Section 6.2.3.4. Hence, CR-SN and CR-MN were selected for training the subjective confidence classifier. For the source memory classifier, remember source responses with correct item memory but different accuracy (SC-RS and SI-RS) were picked for the two classes. For the item memory classifier, responses with correct item memory but lack of source memory (SI) and correct item rejections (CR) were chosen. Finally, the subjective RS vs. F R-K classifier for discriminating between choices in modified R-K judgment were trained with SC-RS and SC&SI-F. SI-RS was excluded from training because the remember source response was not accurate due to the wrong source memory. FA was also excluded from training due to the lack of item or source memory. As a result, four different two-class binary classifiers (CR-SN vs. CR-MN, SC-RS vs. SI-RS, SI vs. CR, and RS vs. F) with projected outputs were trained to discriminate between pairs of

behavioral conditions.

- CR-SN vs. CR-MN classifier

  The CR-SN vs. CR-MN classifier (trained to separate CR-SN from CR-MN, see Figure 6.1) was designed to distinguish different levels of confidence based on presumably no memory strength.

- SC-RS vs. SI-RS classifier

  The SC-RS vs. SI-RS classifier (trained to differentiate between SC-RS and SI-RS, see Figure 6.2) was expected to find a projection which revealed the difference in the amount/accuracy of source information retrieved from the study episode.

- SI vs. CR classifier

  The SI vs. CR classifier (trained to distinguish SI from CR, see Figure 6.3) was proposed to identify the strength of item memory from the study episode without any source memory related effect.

- RS vs. F classifier

  The RS vs. F classifier (trained to discriminate between SC-RS and SC&SI-F, see Figure 6.4) was designed to distinguish remember source from familiar responses in the modified R-K judgment. This is the main judgment that we would like to explain in terms of its simpler component parts.

6.2.3 Methods

A binary classifier using linear discriminant analysis (LDA) with automatic shrinkage regularization as presented in Schäfer and Strimmer (2005) based on Ledoit and Wolf (2004) was trained to classify the feature vectors. In order to avoid overfitting to the training data and take advantage of the consistent spatio-temporal features across subjects, the classifier was trained using leave-one-subject-out (LOSO) cross-validation as given in Liao et al. (2018). The trials

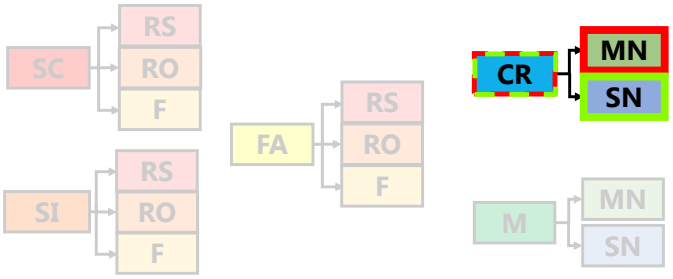


Figure 6.1. CR-SN and CR-MN were selected for training the CR-SN vs. CR-MN confidence classifier. Green box and red box are for the positive and negative class, respectively.

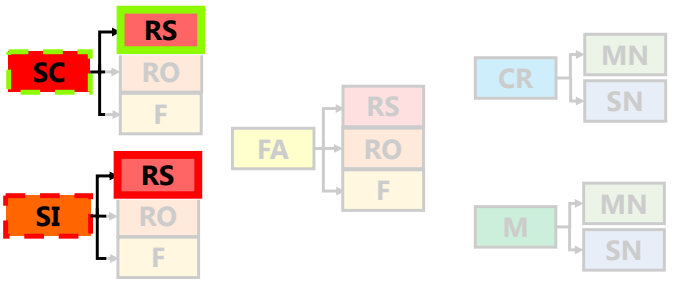


Figure 6.2. SC-RS and SI-RS were selected for training the SC-RS vs. SI-RS source memory classifier. Green box and red box are for the positive and negative class, respectively.

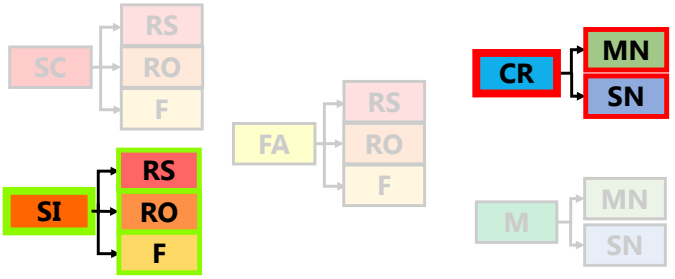


Figure 6.3. SI-RS, SI-RO, SI-F and CR-SN, CR-MN were selected for training the SI vs. CR item memory classifier. Green box and red box are for the positive and negative class, respectively.

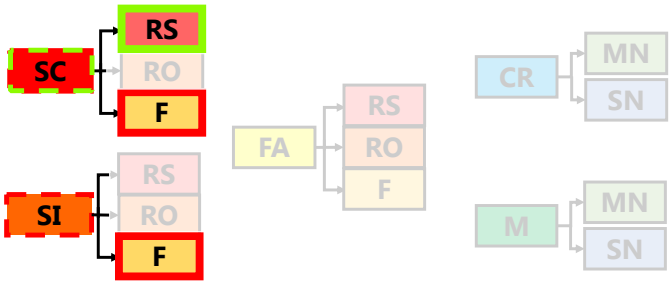


Figure 6.4. SC-RS, SC-F, and SI-F were selected for training the RS vs. F R-K classifier. Green box and red box are for the positive and negative class, respectively.

of the two training classes for each subject remained unbalanced, and these trials from each non-test subject were combined as the LOSO training data. The covariance matrices of two classes were combined, where the numbers of trials were the weights, to obtain a more reliable shared covariance matrix for determining the projection function of the classifier. After training, each classifier was used to project the data onto a 1-dimensional vector which was perpendicular to the classification hyperplane. All the data, including conditions which were not included in the two training classes, of the test subject were projected onto the discriminant vector.

6.2.3.1 Differentiate Conditions not Trained

In addition to classifying the memory strength or confidence of the two training conditions, the classifiers could also identify the difference in memory strength or confidence between any other two conditions of interest (Noh & de Sa, 2014). The selected conditions were paired and projected onto the trained classifiers, and the areas under the ROC curves (AUROCs) for discrimination of the projected two conditions were calculated. The AUROCs of the paired conditions significantly above 0.5 indicated the trained classifier could differentiate the paired conditions and the paired conditions were aligned with the two training classes in the same direction. This suggested the positive condition and the positive training class were more alike. The AUROCs significantly below 0.5 also indicated the trained classifier could differentiate the paired conditions but the paired conditions were in the opposite direction of the training classes. This indicated the tested positive condition and the negative training class were more similar.

The AUROCs not significantly different from 0.5 suggest the two conditions were comparable, and not discriminable, when projected onto the discriminative direction for the trained classifier.

6.2.3.2 Statistical Methods

The activation patterns of each classifier could be obtained as in Haufe et al. (2014). However, due to LOSO cross-validation, a large portion of the training data for classification between any two test subjects were duplicated, which meant consistent spatio-temporal features in the activation patterns of the classifiers between subjects as computed in Haufe et al. (2014) might overestimate consistency between subjects. Alternatively, the mean differences between the spatio-temporal feature vectors of the two training classes for each subject for the paired classification problem in each time window were calculated. The mean difference vector for each subject was L2-normalized to have length 1. In order to identify consistent features across subjects, a cluster-based method for correction for multiple comparisons was then used (Maris & Oostenveld, 2007). In this method, each spatio-temporal pixel (pixel of a channel, time pair) significantly different from zero ($p < .05$) over all subjects was identified. Then the t-statistic of all significant flagged neighboring pixels with the same sign was summed and the maximum absolute value over all clusters taken. This value is compared to the distribution of max absolute cluster values obtained from a permutation distribution resulting from 10,000 random permutations of class labels for each subject. Spatial neighbors were the channel groups containing adjacent electrodes from the cap layout (shown in Figure 2.2). By this definition, LAS, CM, and RAS were all mutual neighbors; CM and PM were neighbors with LPS and RPS. Temporal neighbors were temporally adjacent time windows.

6.2.3.3 Confidence Control (CC)

The AUROCs of SC-RS vs. SI-RS projections and SI vs. CR projections on the CR-SN vs. CR-MN classifier in Experiment 1 were both significantly above 0.5 in the Section 6.2.4. Such phenomena indicate that both the trained SC-RS vs. SI-RS classifier and SI vs. CR classifier

in Experiment 1 could have been trained to perform classification partially based on confidence differences. To control the confidence level of the two training classes in each classifier and keep as many trials for training as possible in a similar way to that in Chapter 5.2.3 and in Liao, Mollison, Curran, and de Sa (2021), we first identified the training class with more trials and calculated the average projections on the CR-SN vs. CR-MN classifier for both training classes. If the average of the training class with more trials was higher than the average of the training class with less trials, we sorted the trials in the first class by their projections onto the CR-SN vs. CR-MN classifier, selected the trials with the smallest values, and accumulated trials in a bottom-up manner until their average was the same as the average of the projections of the trials in the second class in the training data. In contrast, if the average of the training class with more trials was lower than the average of the training class with less trials, we again sorted the trials in the first class, selected the trials with the largest values, and accumulated trials in a top-down manner instead until the their average was the same as the average the second class in the training data. Then, the two training classes with confidence control (as defined by projections from the CR-SN vs. CR-MN classifier) were used as the new training data for training a new SC-RS vs. SI-RS classifier with confidence control (named SC-RS vs. SI-RS (CC)) and a new SI vs. CR classifier with confidence control (named SI vs. CR (CC)).

6.2.3.4 Item Memory Control (ImC)

In Section 6.2.2, we hypothesized that CR-SN vs. CR-MN classifier was substantially classifying based on a confidence difference. However, one could argue that the difference between CR-SN and CR-MN may be more strongly affected by a difference in item memory strength. One intuitive way to address our hypothesis was to compare the CR-SN vs. CR-MN classifier with and without item memory control. To this end, we defined the null hypothesis H_0 and the alternate hypothesis H_1 to be,

- H_0: CR-SN vs. CR-MN classifier is not different from the CR-SN vs. CR-MN classifier with item memory controlled.

- H_1: CR-SN vs. CR-MN classifier and CR-SN vs. CR-MN classifier with item memory controlled are different.

If item memory control for CR-SN vs. CR-MN leads to different results, then the hypothesis should be rejected.

The same procedure as used for confidence control in Sec. 6.2.3.3 can be used for item memory control for the CR-SN vs. CR-MN classifier. The new item memory controlled CR-SN vs. CR-MN classifier could later be used for confidence control for SC-RS vs. SI-RS classifier as shown in Sec. 6.2.3.3.

6.2.4 Results

6.2.4.1 Classifier Performance and Projections

For this analysis, only subjects with more than 5 trials in both of the test classes for each classification problem were considered as test subjects. The accuracy and the AUROC were calculated for each classifier, and the accuracy for each test subject was calculated with balanced classes. The significance of the classification accuracy over chance for each subject was evaluated based on the number of test trials used for classification with small sample size adjustment (Agresti & Caffo, 2000; Müller-Putz, Scherer, Brunner, Leeb, & Pfurtscheller, 2008). The overall average accuracy was also calculated by taking the mean of the test subject accuracy. The significance of the overall average accuracy over chance level was measured using the non-parametric Wilcoxon Signed Rank Test (Wilcoxon, 1945) based on the average accuracy of each test subject, and the significance of the overall AUROC was calculated using the same approach. Table 6.1 and Table 6.2 show the performance of the classifiers in Exp 1, and Table 6.1 and 6.4 present the counterparts of the classifiers in Exp 2. The aggregated AUROCs of each classification problem in each experiment from Table 6.2 and Table 6.4 are also given in Table 6.5 in matched classifier and behavior pair.

The average projection and the 95% confidence interval of each behavior in each classification problem were calculated for both experiment. The order of the behaviors is firstly

Table 6.1. Exp 1 ACCs of test subjects with at least 5 trials in each class.

Classifier Sub. ID	CR-SN vs. CR-MN	SC-RS vs. SI-RS	SC-RS vs. SI-RS (CC)	SI vs. CR	SI vs. CR (CC)	RS vs. F
102	0.5685	0.6000	0.4964	0.5076	0.6031^{*}	0.6382^{**}
103	–	–	–	0.5000	0.4813	0.5821
104	0.5071	0.5188	0.5875	0.6598^{**}	0.6216^{*}	0.6153^{*}
105	–	0.5944	0.5389	$0.6433^{\dagger}$	0.5733	$0.6467^{\dagger}$
106	0.4442	0.5250	0.5278	0.6625^{**}	0.7053^{**}	0.7109^{**}
108	$0.6147^{\dagger}$	0.5500	0.5214	0.5618	0.5745	0.7514^{**}
109	0.5591	–	–	$0.6333^{\dagger}$	0.6048	0.5667
110	0.6825^{*}	–	–	0.4809	0.5221	0.7302^{**}
112	0.4786	0.6917	0.6500	0.6281	0.6000	0.5553
113	0.5537	0.6167	0.5667	0.5568	0.5338	0.7176^{**}
114	0.4769	–	–	0.4845	0.5574	0.7529^{**}
115	0.5364	0.5604	0.5167	0.5667	$0.5757^{\dagger}$	0.7266^{**}
116	0.5071	0.5667	0.4792	0.4740	0.5160	0.5541
117	0.5479	0.5778	0.5583	0.5600	0.5738	0.6487^{**}
118	0.5500	0.7357^{*}	0.4929	0.5324	0.5882	0.6804^{**}
119	$0.5905^{\dagger}$	0.5628	0.5064	0.5458	0.5590	$0.5780^{\dagger}$
120	0.4833	0.5100	$0.7600^{\dagger}$	0.5721	0.5826	0.6531^{**}
121	0.5603	0.5194	0.4889	0.4492	0.4127^{*}	0.6475^{**}
122	0.5405	0.6152	0.6283	$0.5891^{\dagger}$	$0.5913^{\dagger}$	$0.6214^{\dagger}$
123	0.5841	0.5000	0.5286	$0.5939^{\dagger}$	0.6133^{*}	0.5870
124	$0.5921^{\dagger}$	0.6333	0.4994	0.4900	0.5220	0.6042
125	0.5015	–	–	0.5047	0.4906	0.7116^{**}
126	0.5162	–	–	0.4978	0.4772	0.6757^{**}
127	0.5290	–	–	$0.6100^{\dagger}$	$0.6114^{\dagger}$	0.7455^{**}
128	0.5554	–	–	0.5154	0.5256	0.6868^{**}
129	0.5642	0.5433	0.5667	0.4500	0.4323	0.6714^{**}
Avg. ACC	0.5435^{**}	0.5790^{**}	0.5505^{**}	0.5515^{**}	0.5557^{**}	0.6561^{**}

$^{**}\ p < 0.01, ^{*}\ p < 0.05, ^{\dagger}\ p < 0.1$

by the subjective ratings (RS, RO, F, MN, and SN) then by the memory performance within each subjective rating. Figure 6.5 shows the average projection of different behaviors onto the discriminant direction for the trained CR-SN vs. CR-MN confidence classifiers, SC-RS vs. SI-RS source memory classifiers, and SC-RS vs. SI-RS source memory classifiers with confidence control. As for the SI vs. CR item memory classifier, SI vs. CR item memory classifier with confidence control, and RS vs. F R-K classifier, the average projection of the behaviors onto them are presented in Figure 6.6.

In Figure 6.5(1), the CR-SN vs. CR-MN classifier was only trained on correct rejection responses. However, the projection of the SC-RS trials are positive while the projections of

Table 6.2. Exp 1 AUROCs of test subjects with at least 5 trials in each class.

Classifier Sub. ID	CR-SN vs. CR-MN	SC-RS vs. SI-RS	SC-RS vs. SI-RS (CC)	SI vs. CR	SI vs. CR (CC)	RS vs. F
102	0.5781	0.5865	0.5442	0.6302	0.6396	0.6300
103	–	–	–	0.5187	0.4725	0.5509
104	0.5504	0.5482	0.5493	0.6946	0.6912	0.6944
105	–	0.6009	0.5918	0.7042	0.6208	0.7578
106	0.4980	0.5433	0.5767	0.7134	0.7197	0.7805
108	0.7043	0.5745	0.5369	0.6382	0.6355	0.8221
109	0.6321	–	–	0.6847	0.6781	0.7898
110	0.6330	–	–	0.4441	0.4546	0.7826
112	0.5093	0.7833	0.7639	0.6588	0.6571	0.6088
113	0.6029	0.6490	0.6058	0.5838	0.5897	0.7739
114	0.5385	–	–	0.5032	0.5338	0.8563
115	0.5689	0.5578	0.5421	0.6218	0.6332	0.7710
116	0.5789	0.4853	0.4722	0.4882	0.5063	0.5894
117	0.5659	0.6398	0.5833	0.5843	0.5824	0.6967
118	0.5455	0.6401	0.5604	0.5613	0.5326	0.8097
119	0.6484	0.5120	0.4692	0.5412	0.5293	0.6527
120	0.5073	0.7415	0.7262	0.5749	0.5901	0.7260
121	0.5118	0.5618	0.5194	0.4121	0.3895	0.6950
122	0.5499	0.5715	0.5820	0.6346	0.6555	0.5749
123	0.6520	0.4987	0.4610	0.6182	0.6189	0.6405
124	0.6065	0.6667	0.5635	0.5408	0.5403	0.6964
125	0.5211	–	–	0.5255	0.4767	0.7824
126	0.5125	–	–	0.4576	0.4550	0.7074
127	0.5185	–	–	0.6577	0.6677	0.8044
128	0.6493	–	–	0.5191	0.4991	0.7666
129	0.5876	0.5890	0.5633	0.4187	0.4152	0.7109
Avg. AUC	0.5738**	0.5972**	0.5672**	0.5742**	0.5686**	0.7181**

** $p < 0.01$,* $p < 0.05$, † $p < 0.1$

the F responses are negative. Using the CR-SN vs. CR-MN classifier, behaviors having a positive projection means they are more like CR-SN; on the other hand, behaviors with a negative projection mean they are more like CR-MN. Here, the projection of SC-RS indicates it is more like CR-SN (high confidence) rather than CR-MN, and the projections of F show F trials are more similar to CR-MN (low confidence).

In Figure 6.5(2-a), the projections in Exp 1 show a decreasing trend from left to right with projection of SC-RS being the highest. On the other hand, in Figure 6.5(2-b) the projections of SC-RS and SI-RS in Exp 2 are not separated as much as they are in Exp 1. Note that in Figure 6.5(2-a), CR-SN is more positive than CR-MN. If the SC-RS vs. SI-RS classifier performs

Table 6.3. Exp 2 ACCs of test subjects with at least 5 trials in each class.

Classifier Sub. ID	CR-SN vs. CR-MN	SC-RS vs. SI-RS	SC-RS vs. SI-RS (CC)	SI vs. CR	SI vs. CR (CC)	RS vs. F
201	0.4615	0.4171	0.3800*	0.5638	0.5606	0.5167
202	0.5667	0.6800†	0.6450	0.5486	0.5929	0.5553
203	0.5500	0.5200	0.5700	0.6529**	0.6514**	0.7179*
204	0.6200†	–	–	0.5769	0.5808	0.6029
205	0.6167†	0.6190	0.6167	0.5592	0.5690†	0.5588
206	0.5938	0.5234	0.5375	0.6417**	0.6396**	0.5100
207	0.4724	0.6100	0.4900	0.5339	0.5357	0.5053
208	0.5442	0.4756	0.4683	0.5321	0.5236	0.5000
209	0.5455	0.5167	0.6000	0.6937**	0.6635**	0.6844*
210	–	0.5621	0.5629	0.5279	0.5331	–
211	0.5721	0.5265	0.4941	0.5826*	0.5773*	0.5962*
212	0.5056	0.5773	0.5500	0.6092†	0.6105*	0.5767
213	0.5179	0.5364	0.5455	0.5520	0.5740*	0.5250
214	0.5279	0.5333	0.5667	0.5594	0.5775*	0.6583*
215	0.6220†	0.3000†	0.3125†	0.5977**	0.6040**	0.6581*
216	0.6038*	0.4859	0.4650	0.5380	0.5060	0.6342†
217	0.5429	–	–	0.5387	0.5601	0.5875
219	0.5152	0.5214	0.5786	0.5963†	0.6100*	0.5929
220	0.6948**	0.4625	0.4467	0.6492**	0.6517**	–
221	0.6134*	0.5536	0.5191	0.5407	0.5698†	0.4904
222	–	0.5179	0.5429	0.4933	0.4981	0.5000
223	0.5952†	–	–	0.5029	0.5225	0.8162**
224	0.5600	–	–	0.6335**	0.6267**	0.6500†
225	0.6269†	0.6094	0.5875	0.6089*	0.5815†	0.5429
227	0.7643*	0.4435	0.4696	0.5995**	0.5846*	0.5969
228	0.4531	0.5409	0.5591	0.6000†	0.5423	0.5863
229	0.5212	0.5346	0.5628	0.5126	0.5126	0.6755**
230	0.5465	0.4682	0.4409	0.5509	0.5627†	0.5173
Avg. ACC	0.5674**	0.5223†	0.5313†	0.5749**	0.5757**	0.5906**

** $p < 0.01$,* $p < 0.05$, † $p < 0.1$

classification solely based on source memory strength (e.g. source memory accuracy), one should expect the projection of CR-SN to be at the same level or lower than the projection of CR-MN because CR-SN should have no stronger source memory strength than CR-MN . However, the projection of CR-SN is in fact higher than the projection of CR-MN, with a difference that is similar to their difference in the CR-SN vs. CR-MN classifier. This implies that the SC-RS vs. SI-RS classifier is potentially performing classification based on differences in both source memory and confidence in Exp 1. Figure 6.5(2) illustrates the new projections of the responses

Table 6.4. Exp 2 AUROCs of test subjects with at least 5 trials in each class.

Classifier Sub. ID	CR-SN vs. CR-MN	SC-RS vs. SI-RS	SC-RS vs. SI-RS (CC)	SI vs. CR	SI vs. CR (CC)	RS vs. F
201	0.5158	0.3907	0.3989	0.6079	0.6069	0.5628
202	0.6133	0.6148	0.5741	0.5651	0.5476	0.6238
203	0.6156	0.6080	0.5850	0.6476	0.6469	0.7771
204	0.6073	–	–	0.7053	0.7058	0.6516
205	0.5904	0.6636	0.6716	0.6137	0.6184	0.6380
206	0.5827	0.5953	0.5805	0.6667	0.6792	0.5500
207	0.4259	0.7209	0.6744	0.6050	0.6345	0.5288
208	0.5850	0.5077	0.4823	0.5230	0.5085	0.4984
209	0.6265	0.5725	0.5266	0.6890	0.7010	0.7998
210	–	0.6068	0.5858	0.5530	0.5675	–
211	0.6146	0.5169	0.4772	0.6411	0.6397	0.6346
212	0.4789	0.5879	0.5758	0.6555	0.6498	0.6097
213	0.5818	0.5610	0.5755	0.6285	0.6347	0.5418
214	0.6189	0.5604	0.5596	0.6064	0.5990	0.7458
215	0.7161	0.3698	0.3932	0.6595	0.6495	0.6976
216	0.6407	0.4018	0.4036	0.5669	0.5859	0.6876
217	0.5194	–	–	0.5658	0.5725	0.6687
219	0.5433	0.6049	0.6165	0.6424	0.6483	0.5916
220	0.8069	0.4385	0.4349	0.6879	0.6897	–
221	0.7420	0.5462	0.5223	0.5635	0.5662	0.5803
222	–	0.5429	0.5476	0.5075	0.5083	0.4933
223	0.6093	–	–	0.5102	0.5114	0.8884
224	0.5608	–	–	0.6736	0.6853	0.6123
225	0.6186	0.6652	0.6386	0.6714	0.6590	0.6076
227	0.7266	0.5342	0.5485	0.6487	0.6473	0.5607
228	0.4348	0.5591	0.5569	0.5857	0.5849	0.5678
229	0.5691	0.5243	0.5166	0.5293	0.5234	0.7555
230	0.6270	0.4948	0.4918	0.6406	0.6478	0.5451
Avg. AUC	0.6004**	0.5495*	0.5391*	0.6129**	0.6150**	0.6315**

** $p < 0.01$,* $p < 0.05$, † $p < 0.1$

onto the SC-RS vs. SI-RS source memory classifier after confidence control in the training data. Comparing the projections of CR-SN and CR-MN in Figure 6.5(2-a) and (3-a), it should be obvious that the gap between CR-SN and CR-MN was largely reduced after confidence control in the training data for the source memory classifier in Exp 1.

The projections of the responses onto the SI vs. CR item memory classifiers in Exp 1 and 2 are presented in Figure 6.6(4-a) and (4-b), respectively. The projections of the responses in Exp 2 in Figure 6.6(4-a) show again a decreasing trend from left to right for both Exp 1 and 2. The confidence control for training data was also applied to the item classifier to avoid any possible substance of confidence being used for classification. Figure 6.6(5-a) and (5-b) illustrate

Table 6.5. Areas under ROC curves calculated based on the projections of behavior-pairs onto different classifiers

Classifiers \ Behaviors		CR-SN vs. CR-MN	SC-RS vs. SI-RS	SI vs. CR	RS vs. F
CR-SN vs. CR-MN	Exp 1	0.5738**	0.5746*	0.4709*	0.6411**
	Exp 2	0.6004**	0.5250	0.4886	0.6234**
SC-RS vs. SI-RS	Exp 1	0.5576**	0.5972**	0.4860	0.6602**
	Exp 2	0.5485	0.5495*	0.5223*	0.5821**
SC-RS vs. SI-RS (CC)	Exp 1	0.5201	0.5672**	0.5012	0.6097**
	Exp 2	0.5360	0.5391*	0.5168*	0.5572**
SI vs. CR	Exp 1	0.4632	0.4750	0.5742**	0.5008
	Exp 2	0.4654	0.4929	0.6129**	0.5860**
SI vs. CR (CC)	Exp 1	0.4721	0.4938	0.5686**	0.5339
	Exp 2	0.4675	0.4950	0.6150**	0.5875**
RS vs. F	Exp 1	0.5670**	0.6167**	0.4750	0.7181**
	Exp 2	0.5973**	0.5153	0.5577**	0.6315**

** $p < 0.01$,* $p < 0.05$

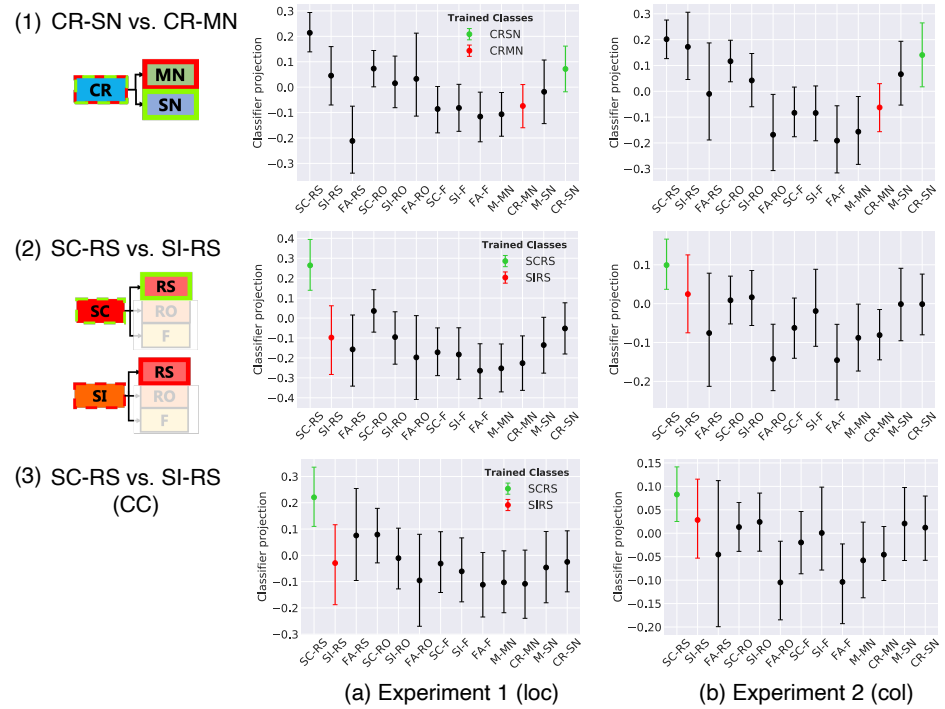


Figure 6.5. The average projection and the 95% confidence interval of behaviors from (1) CR-SN vs. CR-MN confidence classifier, (2) SC-RS vs. SI-RS source memory classifier, and (3) SC-RS vs. SI-RS source memory classifier with confidence control in (a) Exp 1 and (b) Exp 2. The positive and negative trained classes are plotted in green and red, respectively. The classes shown in black were not trained.

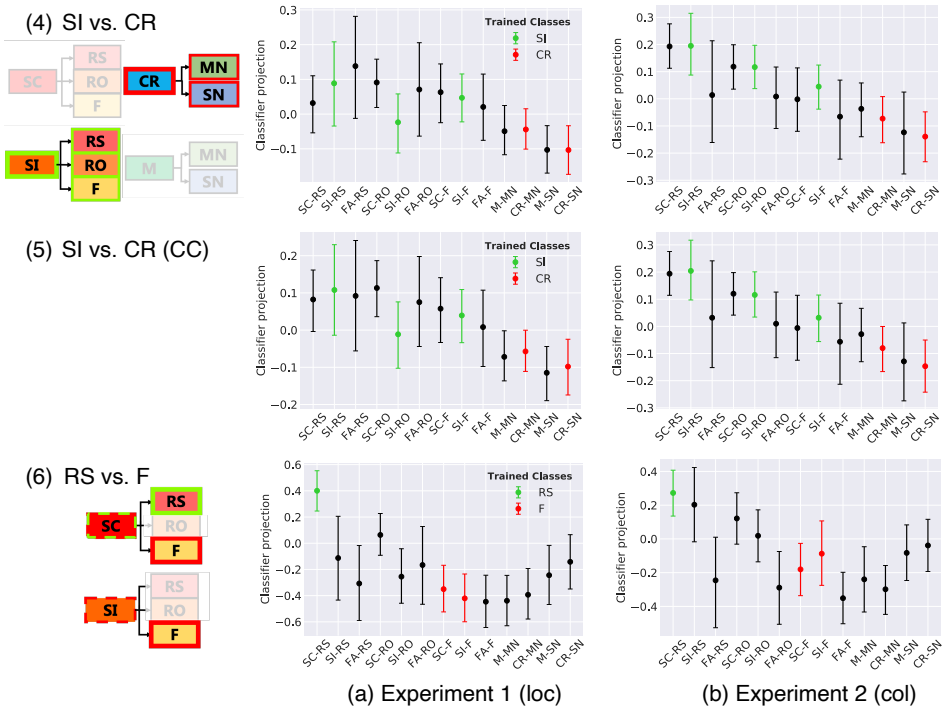


Figure 6.6. The average projection and the 95% confidence interval of behaviors from (4) SI vs. CR item memory classifier, (5) SI vs. CR item memory classifier with confidence control, and (6) RS vs. F R-K classifier in (a) Exp 1 and (b) Exp 2. The positive and negative trained classes are plotted in green and red, respectively. The classes shown in black were not trained.

the projections of the behaviors onto the SI vs. CR classifiers with confidence control for the training data in Exp 1 and 2, respectively.

In Figure 6.6(6), the projections of the old responses from the RS vs. F classifier on the left half again show a decreasing trend from left to right like the counterparts in Figure 6.5(1) for CR-SN vs. CR-MN classifier, Figure 6.5(3) for SC-RS vs. SI-RS classifier with confidence controlled, and Figure 6.6(5) for SI vs. CR classifier with confidence controlled. On the contrary, like in Figure 6.5(1), the projections of the new responses from the RS vs. F classifier in Figure 6.6(6) show an increasing trend from left to right. In fact, the projections in Figure 6.6(6) appear to be approximately the combination of the projections from CR-SN vs. CR-MN classifier in Figure 6.5(1), the projections from SC-RS vs. SI-RS classifier with confidence controlled in Figure 6.5(3), and the projections from SI vs. CR classifier with confidence controlled in

Figure 6.6(5).

6.2.4.2 Patterns for Classification

The mean differences of the extracted spatio-temporal features between the two classes for each classification problem along with the significance of the features across subjects for Exp 1 and 2 are presented in Figure 6.7 and Figure 6.8.

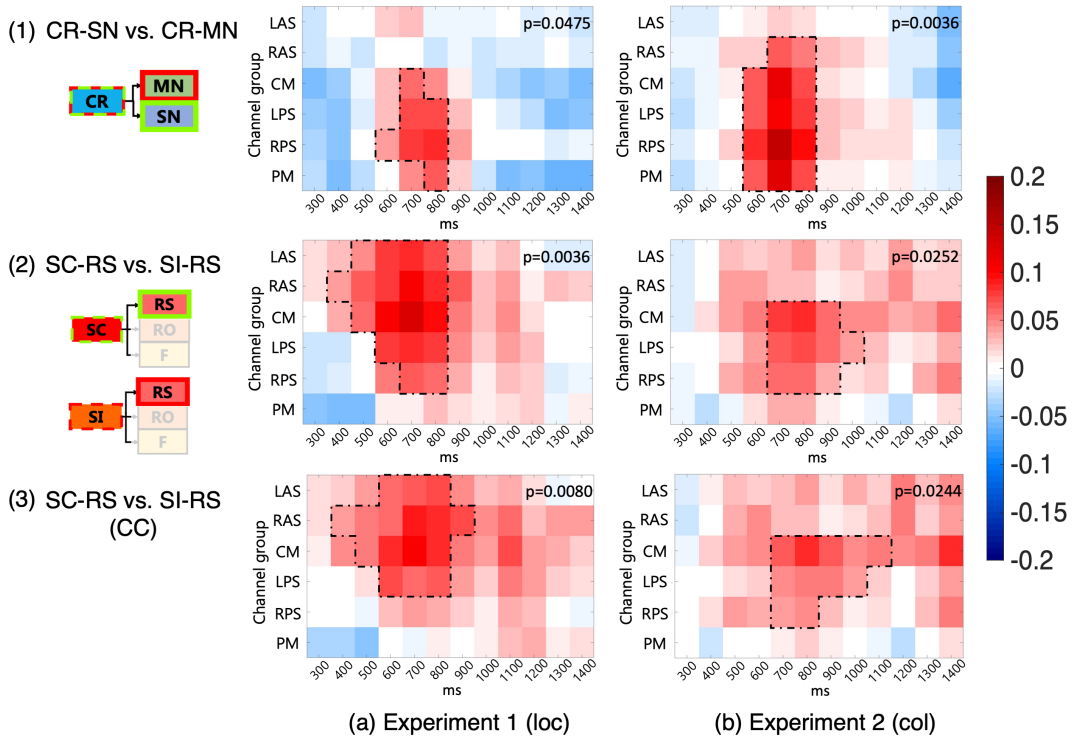


Figure 6.7. Mean difference between training classes of each classification problem (1: CR-SN vs. CR-MN, 2: SC-RS vs. SI-RS, 3: SC-RS vs. SI-RS with confidence control) in (a) Experiment 1 and (b) Experiment 2. The areas surrounded by the dotted lines are the top positive/negative clusters with p-value $< .05$. The p-value on the bottom right of each panel is calculated based on the cluster with the largest absolute t-stat.

The significant clusters in the CR-SN vs. CR-MN confidence classification problem for Exp 1 and 2 in Figure 6.7(1) both peak around 700-800 ms and are mostly distributed at parietal, which indicates the underlying cognitive process for decision confidence could be similar regardless of the source type in the experiment. In the SC-RS vs. SI-RS source memory classification problem without confidence control, the significant cluster is broader for

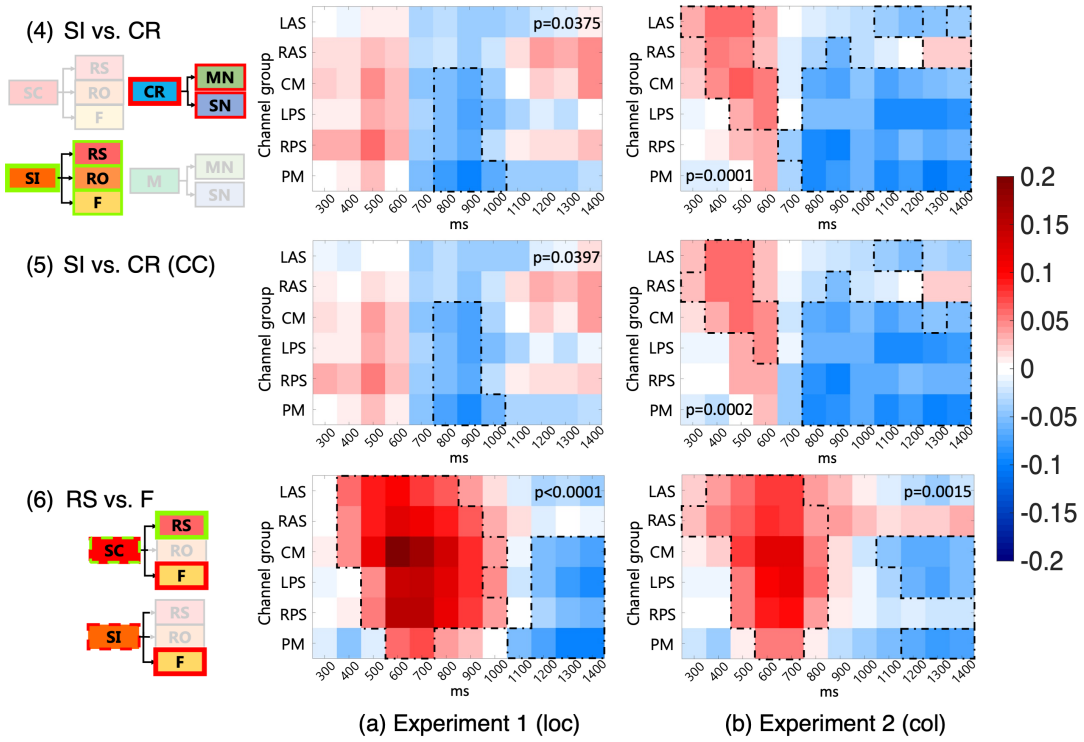


Figure 6.8. Mean difference between training classes of each classification problem (4: SI vs. CR, 5: SI vs. CR with confidence control, 6: RS vs. F) in (a) Experiment 1 and (b) Experiment 2. The areas surrounded by the dotted lines are the top positive/negative clusters with p-value $< .05$. The p-value on the bottom right of each panel is calculated based on the cluster with the largest absolute t-stat.

location source (e.g. Exp 1, see Figure 6.7(2-a)) than for the frame color source (e.g. Exp 2, see Figure 6.7(2-b)). With confidence control, the significant cluster in Exp 1 becomes more center- and left-parietal distributed as shown in Figure 6.7(3-a).

The patterns for the SI vs. CR item memory classification problem are fairly similar with confidence control in Figure 6.8(4) and without confidence control in Figure 6.8(5). The early positive components appear both in location source and frame color source experiment, but only the one in the frame color source experiment reveals significance. In addition, the early component is more frontally distributed in the frame color source experiment than in the location source experiment. In the RS vs. F r-k classification problem, the patterns in Figure 6.8(6) for both Exp 1 and 2 are similar. The positive clusters both peak at 600-700 ms and heavily distributed in the parietal region.

6.2.4.3 Projections of Trials Using Different Projection Functions

Table 6.5 shows the AUROCs of the different behavior-pairs behaviors in each classification problem. The significance of AUROC being above 0.5 was calculated in the same way described for overall significance across subjects in Section 6.2.4.1. The blocks with the same behavior-pair for both the row and the column show the performance of each classifier, and all the AUROCs in these blocks are significantly above 0.5. This suggests the performance of the trained classifiers is significantly above chance level. The performance of the CR-SN vs. CR-MN classifier is better in Exp 2 than in Exp 1, which corresponds to the wider significant cluster in Figure 6.7(1-b) than in (1-a). Likewise, the higher AUROCs for SC-RS vs. SI-RS source memory classifier with and without confidence control in Exp 1 than in Exp 2 are associated with the broader significant clusters in Exp 1 in Figure 6.7(2,3-a).

Besides using the classifier with the identical training conditions, each behavior-pair could also be classified by some other classifiers. The CR-SN vs. CR-MN confidence classifier could distinguish SC-RS from SI-RS with AUROC significantly above 0.5 in Exp 1, and SC-RS vs. SI-RS source memory classifier without confidence control could differentiate CR-SN and CR-MN in Exp1. These suggest that there could be a confidence difference between SC-RS and SI-RS, and SC-RS reflects higher confidence than SI-RS in Exp 1 as shown in Figure 6.5(1-a). The SI vs CR difference also appears to include a component of confidence in Exp 1 because they could be differentiated by the CR-SN vs. CR-MN confidence classifier where SI reflects lower confidence than CR.

With confidence control, the performance of the SC-RS vs. SI-RS source memory classifier drops a little in both Exp 1 and Exp 2. However, the SC-RS vs. SI-RS classifier with confidence control could no longer separate CR-SN from CR-MN as judged by having an AUROC close to 0.5 in Exp 1. This implies that the source memory classifier with confidence control is not using the underlying confidence difference between SC-RS and SI-RS in Exp 1. For SI vs. CR item memory classifiers they could not successfully differentiate CR-SN and

CR-MN either with or without confidence control in both experiments.

The fact that RS vs F could be classified by the CR-SN vs. CR-MN confidence classifier in both experiments indicates that the difference in confidence in old responses and the difference between R and K responses have very similar spatio-temporal EEG scalp responses as illustrated in Figure 6.8(6) and Figure 6.7(1). In addition, RS and F could also be differentiated by the source memory classifier with confidence control in both experiments, which corresponds to the overlap of the positive clusters in Figure 6.7(3) and Figure 6.8(6). The SI vs. CR item memory classifier with confidence control could separate RS and F in the frame color source experiment but not in the location source experiment. Significant clusters in Figure 6.8(5) also show more overlap in early frontal and late posterior areas within the significant clusters in Figure 6.8(6) in Exp 2 than Exp 1.

6.2.4.4 Patterns and Projections of CR-SN vs. CR-MN Classifier with and without Item Memory Controlled

The average AUROCs of the projections of different behavior-pairs onto the CR-SN vs. CR-MN classifiers before and after item memory controlled are shown in Table 6.6. To test our hypothesis that item memory is not the major component in terms of the difference between the CR-SN and CR-MN, Wilcoxon rank sum test was performed on users' AUROCs to test if item memory control would lead to the rejection of the null hypothesis which was that the CR-SN vs. CR-MN classifier was not different from the CR-SN vs. CR-MN classifier with item memory controlled (defined in Section 6.2.3.4), and the results are shown in Table 6.7. In every behavior-pair, there is no evidence showing that performing item memory control would fundamentally change the nature of the CR-SN vs. CR-MN classifier.

The averages of projection of each behavior from the CR-SN vs. CR-MN classifier with item memory controlled across those subjects with more than 5 trials for the behavior in Exp 1 and Exp 2 are shown in Figure 6.9. While projections of the behaviors in Exp 1 in Figure 6.9 (a) and (c) share very similar distributions, the scatter plots in Figure 6.9 (e) and the regression line

Table 6.6. Areas under ROC curves calculated based on the projections of behavior-pairs onto different classifiers

Classifiers \ Behaviors		CR-SN vs. CR-MN	SC-RS vs. SI-RS	SI vs. CR	RS vs. F
CR-SN vs. CR-MN	Exp 1	0.5738^{**}	0.5746^{*}	0.4709^{*}	0.6411^{**}
	Exp 2	0.6004^{**}	0.5250	0.4886	0.6234^{**}
CR-SN vs. CR-MN (ImC)	Exp 1	0.5578^{**}	0.5612^{**}	0.4983	0.6341^{**}
	Exp 2	0.6025^{**}	0.5333	0.5156	0.6432^{**}

** $p < 0.01$,* $p < 0.05$

Table 6.7. Wilcoxon rank sum test on the AUROCs of subjects from CR-SN vs. CR-MN classifier before and after item memory control.

Stats \ Behaviors		CR-SN vs. CR-MN	SC-RS vs. SI-RS	SI vs. CR	RS vs. F
Exp 1	Z-stat	0.5766	0.3807	−1.3085	0.2196
	p-value	0.5642	0.7034	0.1907	0.8262
Exp 2	Z-stat	0	−0.1723	−1.2044	0.6146
	p-value	1	0.8632	0.2284	0.5388

further reveal the high correlation between the Figure 6.9 (a) and (c). The same relationship for Exp 2 could also be found in Figure 6.9 (b), (d), and (f).

The differences in patterns of the CR-SN vs. CR-MN classifier and the CR-SN vs. CR-MN classifier for Exp 1 and Exp 2 are presented in Figure 6.10 (e) and (f), respectively. For Exp 1, there is not any significant cluster in the pattern in Figure 6.10 (e), meaning the pattern of the CR-SN vs. CR-MN classifier before and after item memory control does not really change. As for Figure 6.10 (f) for Exp 2 , there is a positive cluster showing the difference in the classifier before and after item memory control. However, this positive cluster does not overlap with the significant cluster for either the CR-SN vs. CR-MN classifier or the CR-SN vs. CR-MN classifier with item memory controlled. This again indicates that performing item memory control or not does not change the nature of the CR-SN vs. CR-MN classifier.

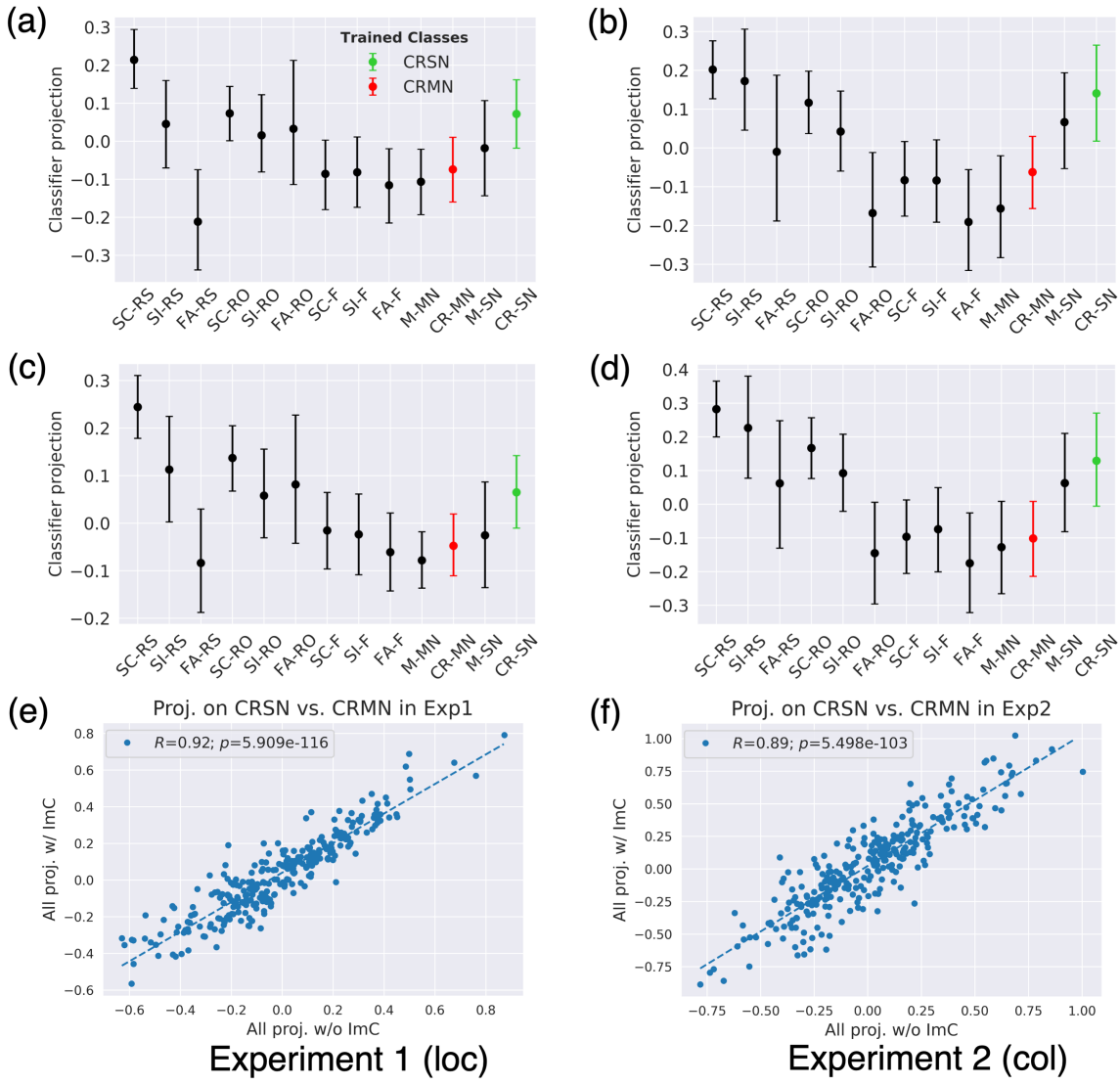


Figure 6.9. (a) and (b) are the average projection and the 95% confidence interval of behaviors from CR-SN vs. CR-MN classifier in Exp 1 and Exp 2, respectively (same as Figure 6.5 (a) and (b)). (c) and (d) are counterparts from CR-SN vs. CR-MN with item memory control in Exp 1 and Exp 2, respectively. (e) and (f) are the scatter plots where the x value and the y value of each dot are the average projection of each behavior from each subject from the CR-SN vs. CR-MN classifier and from the CR-SN vs. CR-MN classifier with item memory controlled, respectively.

6.2.5 Discussion

6.2.5.1 Confidence Classifier

To train a confidence classifier without any memory effect (i.e. familiar, recollection) involved, we deliberately selected only the correct rejection responses to train the confidence

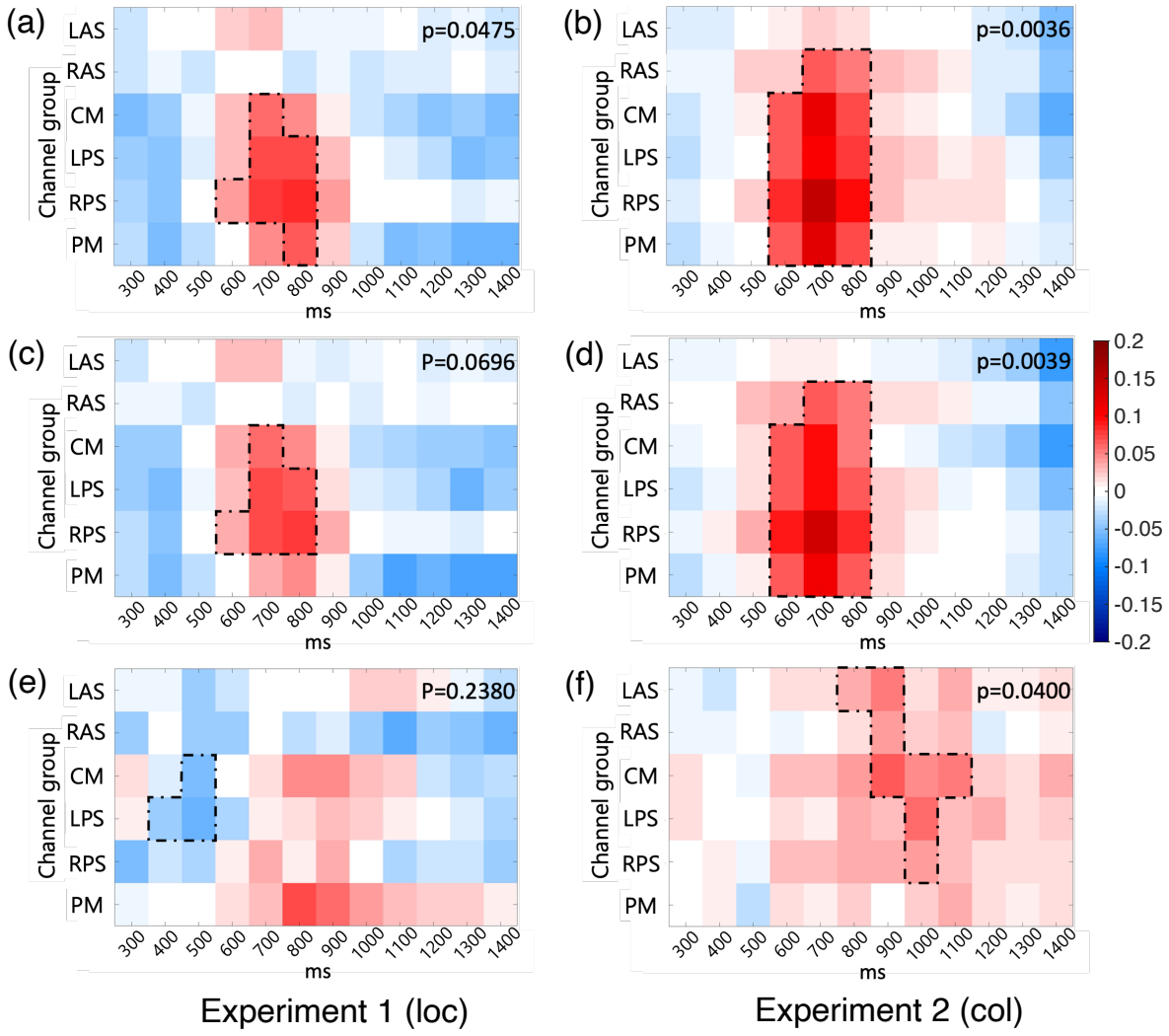


Figure 6.10. (a) and (b) are the mean differences of CR-SN vs. CR-MN in Exp 1 and Exp 2, respectively (same as Figure 6.7 (a) and (b)). (c) and (d) are the mean differences of CR-SN vs. CR-MN with item memory control in Exp 1 and Exp 2, respectively. (e) and (f) are the mean differences between CR-SN vs. CR-MN without and with item memory control in Exp 1 and Exp 2, respectively. The areas surrounded by the dotted lines are the top positive/negative clusters with p-value $< .05$. The p-value on the bottom right of each panel is calculated based on the cluster with the largest absolute t-stat.

classifier.

In Table 6.5, the projections of RS and F from CR-SN vs. CR-MN classifier show that the relationship (or alignment) of RS and F are similar to that of CR-SN and CR-MN because the AUROCs are significantly higher than 0.5. If the CR-SN vs. CR-MN classifier performed

the classification based primarily on item memory or source memory strength, the AUROCs of the projections of RS and F should at or lower than 0.5 since CR-SN should reflect lower item memory or source memory than CR-MN. However, the AUROCs for both Exp 1 and Exp 2 are much higher than 0.5. Such a finding reveals that the main underlying component of the classifier is not memory related, and the best interpretation for this would be the CR-SN vs. CR-MN classifier is much more of a confidence classifier than a memory classifier.

Furthermore, in Section 6.2.4.4, we showed that CR-SN vs. CR-MN classifier was mostly based on the difference in confidence levels instead of item memory because its property did not really change when it came to item memory control.

The results in Section 6.2.4.1 show that it is possible to discriminate the confidence level in memory based on subject-independent training and single-trial scalp EEG activity recorded during the recognition phase in a memory task based on the findings on CR-SN vs. CR-MN confidence classifier. Moreover, the difference in confidence levels between the behavior-pair could be separated by CR-SN vs. CR-MN confidence classifier. The confidence effect in frame color source memory tasks is stronger than in location source memory tasks. In fact, the confidence effect varied by the source type could correspond to the difference in confidence ERPs in Wynn et al. (2019); Wynn, Kessels, and Schutter (2020) led by two different types of stimuli.

6.2.5.2 Source Memory and Item Memory Classifier with Confidence Control

When considering the source accuracy of the RS responses in Table 6.8 and the AUROCs of SC-RS vs. SI-RS from CR-SN vs. CR-MN confidence classifier in Table 6.5, Exp 1 reveals a higher source accuracy and stronger correlation between SC-RS vs. SI-RS and confidence while Exp 2 only shows weak to zero correlation between source memory and confidence , which implies that the source type has an effect on the source accuracy and the correlation between SC-RS vs. SI-RS and confidence (Busey, Tunnicliff, Loftus, & Loftus, 2000). As a result, the trained SC-RS vs. SI-RS source memory classifier in Exp 1, which partly performs classification

using the difference in confidence, could even distinguish CR-SN from CR-MN where source memory difference is absent. On the other hand, the trained SC-RS vs. SI-RS source memory classifiers with confidence control in both Exp 1 and 2 could still distinguish SC-RS from SI-RS but no longer disriminate CR-SN from CR-MN. Such source memory classifiers with confidence control depend more on the source memory difference than the classifiers without confidence control. Contrary to the source memory classifier, the confidence effect does not have as much impact on SI vs. CR item memory classifier as on the source memory classifier.

Table 6.8. Total numbers of trials of correct item responses in both experiments.

# of trials	SC-RS	SI-RS	SC-RO	SI-RO	SC-F	SI-F	CR-MN	CR-SN
Exp 1	2972	276	888	409	616	484	1043	1560
Exp 2	2089	584	1295	1057	488	389	771	1678

	SC/SC&SI	SC-RS/SC&SI-RS
Exp 1	0.7931	0.9150
Exp 2	0.6560	0.7815

6.2.5.3 Patterns for Classification Problems

Unlike the FN400 effect shown in Curran (2004); Wynn et al. (2019) for different confidence levels in new responses, the patterns of CR-SN vs. CR-MN showed there was no significant difference in 250-550 ms in either Exp 1 or 2 in Figure 6.7(1). On the other hand, the LPC effect, where the average of normalized voltage of CR-SN is greater than CR-MN, shown in this chapter is consistent with the findings for the difference in confidence (Curran, 2004; Wynn et al., 2019, 2020). The LPN was not significant for SN vs. MN, which is consistent with the ERP studies (Johansson & Mecklinger, 2003; Mecklinger, 2006) that refer to the LPN as reflecting source reconstruction. Source reconstruction should not be present in either CR-SN or CR-MN.

In the patterns of SC-RS vs. SI-RS in Figure 6.7(2), the difference is stronger for Exp 1, showing that the source memory effect is stronger for the location source experiment. Such a difference in strength is in accordance with the higher percentage of SC trials from the total

number of correct item responses as shown in Table 6.8, (or better source accuracy) in Exp 1 than in Exp 2. While patterns in both Exp 1 and 2 have spatio-temporal features overlapped with LPC effect, the pattern for Exp 1 in Figure 6.7(3-a) becomes more frontally distributed, and the pattern for Exp 2 in Figure 6.7(3-b) shrinks with confidence control. The transformation with confidence control in both experiments again suggests that the LPC is associated with confidence. The pattern of SC-RS vs. SI-RS with confidence control in Exp 1 shows little FN400 effect (Curran, 2000; Mecklinger, 2006; Rugg & Curran, 2007), indicating the familiarity could potentially contribute to source accuracy in the location source experiment. The LPN effect is absent in the pattern of SC-RS vs. SI-RS for both experiment, which matches the previous ERP studies that LPN is not varied with source accuracy (Johansson & Mecklinger, 2003; Herron, 2007).

The patterns of SI vs. CR for both experiment are not considerably changed with confidence control, implying that the average confidence levels for the correct old responses when source memory is removed and for the correct new responses are very close. The FN400 effect is very clear in the pattern of SI vs. CR for Exp 2 but not for Exp 1. This shows that familiarity is relatively more important to item memory when the type of source information is more difficult to recollect, e.g. frame color against item location. The onset of the negative clusters in the patterns of SI vs. CR for both experiment is later than the onset of the LPC. Previous studies (Mollison & Curran, 2012)(Woroch & Gonsalves, 2010) selected the window representing the LPC effect, hence the the positivity from the LPC in the first half of the selected window and the negativity from the LPN in the second half of the selected window cancelled each other out and only showed null effect in the ERP differences. The fact that this negative cluster is mostly distributed in parietal and comes in slightly later than LPC matches the traits of the LPN effect very well. Further studies are required to identify the potential connection between item memory and the LPN effect.

In the patterns of RS vs. F in Figure 6.8(3), the parietal old/new effect appears as previously shown in episodic memory studies (Curran, 2004; Vilberg et al., 2006; Woroch & Gonsalves, 2010), and the spatio-temporal features of parietal old/new effect highly overlap with

the ones in the significant cluster in CR-SN vs. CR-MN patterns in both experiments. For the FN400 effect in the patterns of RS vs. F, the onset in Exp 1 is consistent with the counterpart in the pattern of SC-RS vs. SI-RS in Exp 1, and the distribution in Exp 2 accords with the one in the pattern of SI vs. CR in Exp 2. The LPN effect between RS and F could also be observed after 900 ms at posterior site, corresponding to the ERP study during source monitoring (Leynes & Phillips, 2008). The LPN effect observed in RS vs. F is consistent with the argument that LPN reflects an additional inspection of retrieved information for feature conjunctions (Johansson & Mecklinger, 2003) because RS judgments were made with perceived successful reconstruction of the prior episode.

6.2.5.4 Projection Reveals Components Attribute to R-K Difference

Remember responses associated with relatively high confidence conditions (Dunn, 2004; Yonelinas et al., 2002; Wixted, 2007). However, it is difficult to disassociate memory processes and confidence when investigating the difference between remember responses and familiarity responses in ERP studies. The CR-SN vs. CR-MN classifier was deliberately trained to distinguish the confidence level with minimal effects of memory strength, and it yielded the opportunity to identify the confidence independently in R-K experiments. In addition, by using the projections of the training data onto the confidence classifier, the source memory and item memory classifiers with confidence control could be utilized to decompose the intertwined memory effects between remember and familiarity responses without the complication from confidence.

For the location source Experiment 1 in Table 6.5, RS and F could be differentiated by the CR-SN vs. CR-MN confidence classifier and the SC-RS vs. SI-RS source memory classifier with confidence control. However, they could not be separated by the SI vs. CR item memory classifier with confidence control. Since the performance of the item memory classifier with confidence control is significantly above chance level, the only explanation for why RS and F could not be separated is that the item memory is not a substantial component in the difference

between RS and F.

For the frame color source Experiment 2, the confidence classifier, source memory and item memory classifiers with confidence control could all separate RS and F successfully. The results suggest that the difference between RS and F in the frame color source experiment could be accounted for by confidence, source memory, as well as item memory.

6.3 Interpretation of RS vs. F using Memory and Confidence Components

The AUROCs in Table 6.5 showed that the confidence difference was the same for both old and new judgments in episodic memory retrieval and the difference between RS and F is associated with source memory, item memory, and confidence. However, it remains unclear how each factor contributes to the aggregated difference. In this section, we would like to further decompose the representation of RS vs F in terms of source memory, item memory, and confidence level. To this end, we combined the projections onto the source memory classifier and item memory classifier as well as the projection onto the confidence classifier trained only on new judgments to predict the projection onto the RS vs. F classifier. The projection of the SC-RS vs. SI-RS classifier with confidence control was considered as the index of the source memory strength; the projection of the SI vs. CR classifier with confidence control was considered as the index of the item memory strength; the projection of the CR-SN vs. CR-MN classifier was considered as the index of confidence level.

6.3.1 Methods

6.3.1.1 Linear Regression Model to Predict RS and F

The training process of the CR-SN vs. CR-MN, SC-RS vs. SI-RS with confidence control, and SI vs. CR with confidence control classifiers was the same LOSO paradigm as described in Section 6.2.3, and the selected behaviors for training in each class in each classifier were the same as in Section 6.2.2. After training the classifiers, all the trials in the training

subjects were projected onto the classifiers, and projections of the SC-RS and SC&SI-F trials were selected and balanced to train the linear regression model. Each trial could be viewed as a training sample, the projections onto the confidence, source memory, and item memory classifier were considered as predictor variables, and the projection onto the RS vs. F classifier was the target value. Training data were normalized in order to have the coefficients in the model reflect the dominance of the three components. The trained model was then validated with the projections onto the three classifiers of the trials in the left-out subject with at least 5 trials in both SC-RS and SC&SI-F conditions.

6.3.1.2 Importance of Source Memory and Item Memory in Linear Regression

In Sec. 6.2.4, we have observed that the performance of the source memory and item memory classifiers and the sizes of the significant clusters in the patterns were remarkably different depending on the source type (e.g. different in Exp 1 and Exp 2). As a result, we would like to investigate the importance of the various components by an ablation analysis where we train the linear regression models using confidence and either source memory or item memory as two variables and compare the performance of the model using confidence with both source and item memory to the model using confidence with either source or item memory. The model with confidence, source memory, and item memory as three variables is named CoSmIm model by taking the initial of each variable; the model with confidence and source memory as two variables is called CoSm model, whereas the model with confidence and item memory is called CoIm model.

6.3.2 Results

6.3.2.1 Performance of Trained Classifiers and Linear Regression Models

Table 6.9 presents the performance of different classifiers and linear regression models classifying RS and F with the AUROCs in Exp 1 and 2. Leave-one-trial-out from each class (LOTO) classifier (Noh et al., 2018), which was individual dependent and the most widely used

Table 6.9. The average AUROCs of classifying SC-RS and SC&SI-F using different classifiers and linear regression models.

AUROCs of RS vs. F	LOSO Classifier	CoSmIm Linreg	LOTO Classifier	CoSm Linreg	CoIm Linreg	SmIm Linreg
Exp 1	0.7181^{**}	0.6725^{**}	0.6803^{**}	0.6722^{**}	0.6381^{**}	0.6106^{**}
Exp 2	0.6315^{**}	0.6558^{**}	0.6137^{**}	0.6237^{**}	0.6546^{**}	0.6049^{**}

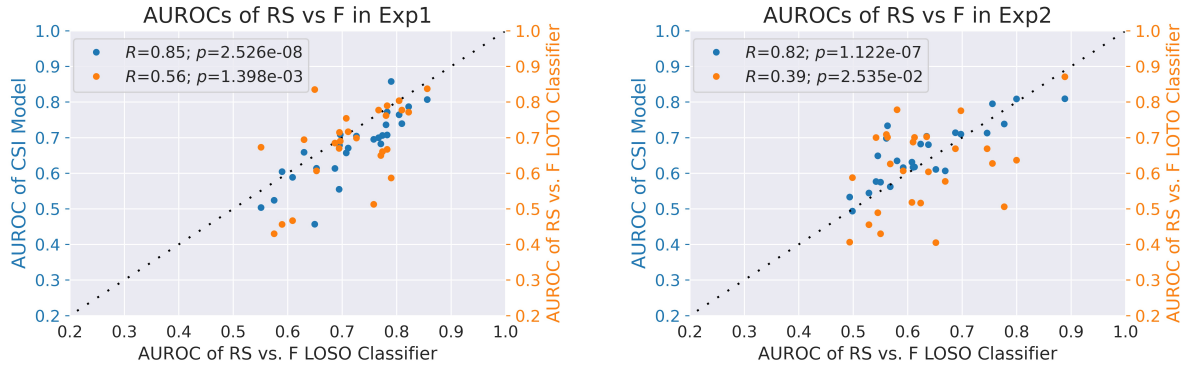
** $p < 0.01$,* $p < 0.05$, † $p < 0.1$

type of classifier for single-trial studies, was also trained for each subject and compared. In both Exp 1 and 2, LOSO classifiers outperform LOTO classifiers. While the AUROCs from the LOSO classifier and from the CoSmIm models are comparable, the LOSO classifier works significantly better than the CoSmIm model in Exp 1, whereas the CoSmIm model performs significantly better than the LOSO classifier in Exp 2. Figure 6.11 illustrates the AUROCs of RS vs. F LOSO classifier, CoSmIm model, and RS vs. F LOTO classifier of each subject. The correlations between AUROCs of RS vs. F LOSO classifier and CoSmIm model are highly significant in both Exp 1 and 2, implying that the CoSmIm model is a decent approximation of the RS vs. F LOSO classifier.

Models with three variables (e.g., CoSmIm) consistently perform better than models with only two variables (e.g. CoSm, CoIm, and SmIm). Table 6.10 shows the p-values of the paired one-tailed t-test of the AUROCs of each test subject using the selected models. In Exp 1, the AUROC of the CoSmIm model is significantly greater than the CoIm model but not than the CoSm model, meaning the source memory is a important variable to the CoSmIm model in differentiating RS from F in location source experiment. In Exp 2, the AUROC of CoSmIm model is greater than the CoSm model but not the CoIm model, indicating the item memory is a substantial factor to the CoSmIm model in separating RS and F in frame color source experiment.

6.3.2.2 Coefficients in Linear Regression Models

Table 6.11 shows the coefficient and the ratio of each predictor variable in the different models. The intercepts of all models trained are relatively tiny, which implies the combined



(a) Experiment 1 (loc) (b) Experiment 2 (col)

Figure 6.11. (a) The AUROCs of SC-RS vs. SC&SI-F projection on RS vs. F LOSO classifier (along x-axis), predictions of CoSmIm linear regression model using classifier (along left y-axis), and projection on RS vs. F LOTO classifier (along right y-axis) of each subject in Exp 1 and (b) Exp 2. The R and p-values indicate the correlation the significance of the correlation between AUROCs in x- and y-axis.

Table 6.10. The p-values of the paired one-tailed Student's t-test between subject AUROCs of different linear regression models.

CoSmIm AUROC	Experiment 1			Experiment 2		
	vs. CoSm	vs. CoIm	vs. SmIm	vs. CoSm	vs. CoIm	vs. SmIm
p-value	0.2511	0.0071	1.067e-5	0.0063	0.3275	0.0011

mean of the predictor variables is close enough to the mean of the target in each model. In the CoSmIm models, the proportion of the item memory term is a lot smaller than the ratio of the source memory term for Exp 1, whereas the ratio of the item memory term is larger than the ratio of the source memory term. This again implies the difference between RS and F in location source experiment could be ascribed to source memory more than item memory; on the other hand, the difference between RS and F in frame color source experiment could be attributed to item memory more than source memory as mentioned above.

6.3.2.3 Projections and Patterns of the CoSmIm Model

The CoSmIm model could also be applied to predict the projections onto RS vs. F classifiers for responses not used for training the CoSmIm model. The mean and 95% con-

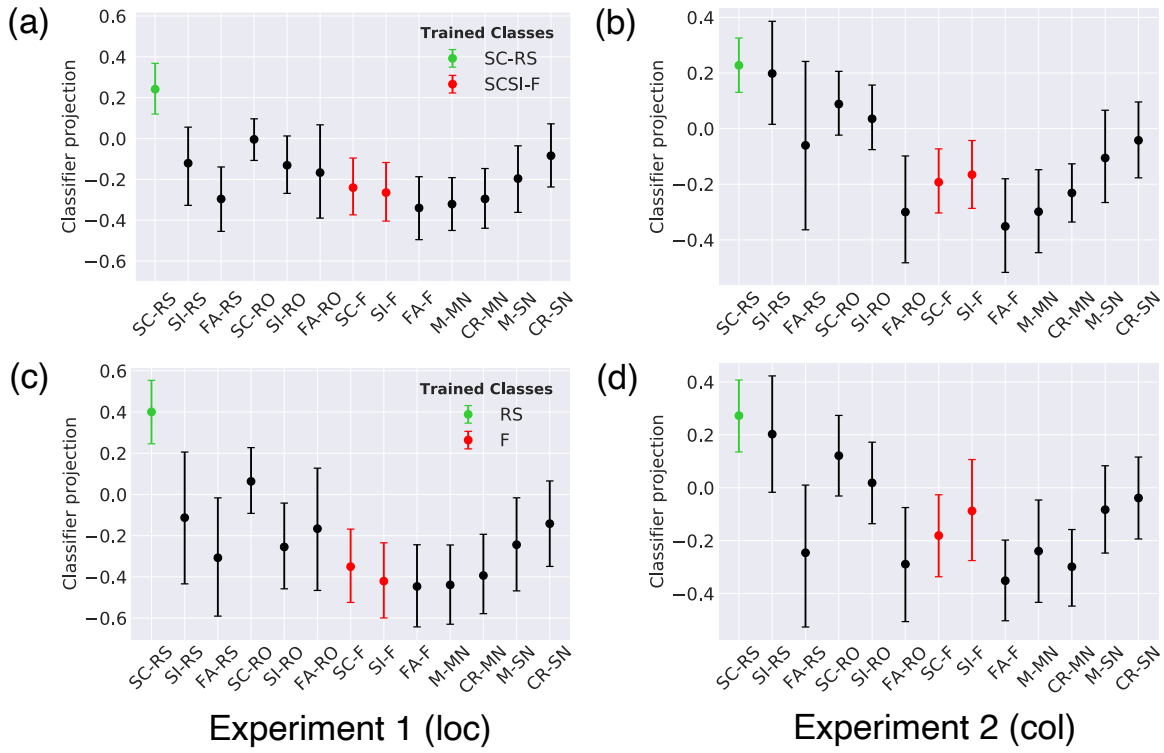


Figure 6.12. (a) The average projection and 95% confidence interval of different behaviors from CoSmIm model for RS and F in Exp 1 and (b) Exp 2. (c) The average projection of different behaviors from RS vs. F R-K classifier in Exp 1 (same as Figure 6.6(6-a)) and (d) Exp 2 (same as Figure 6.6(6-b)) Behaviors in green were trained as the positive class, and behaviors in red were trained as the negative class.

fidence intervals of the predicted projections of all responses in Exp 1 and 2 are illustrated in Figure 6.12(a) and (b), respectively. The high similarity between Figure 6.12(a), (b) and Figure 6.12(c), (d) implies how well the CoSmIm model models the difference between RS and F in held out subjects.

The pattern for the CoSmIm model results from the combination of patterns of CR-SN vs. CR-MN, SC-RS vs. SI-RS with confidence control, and SI vs. CR with confidence control, where the ratio of each term in the model was the corresponding weights for combination. The pattern of CoSmIm for each test subject was generated, and the mean pattern and the significant cluster were calculated in the same manner in Section 6.2.3.2. Although the positive significant clusters in Figure 6.13(a) and (b) are not as broad as the significant positive clusters in Figure 6.13(c)

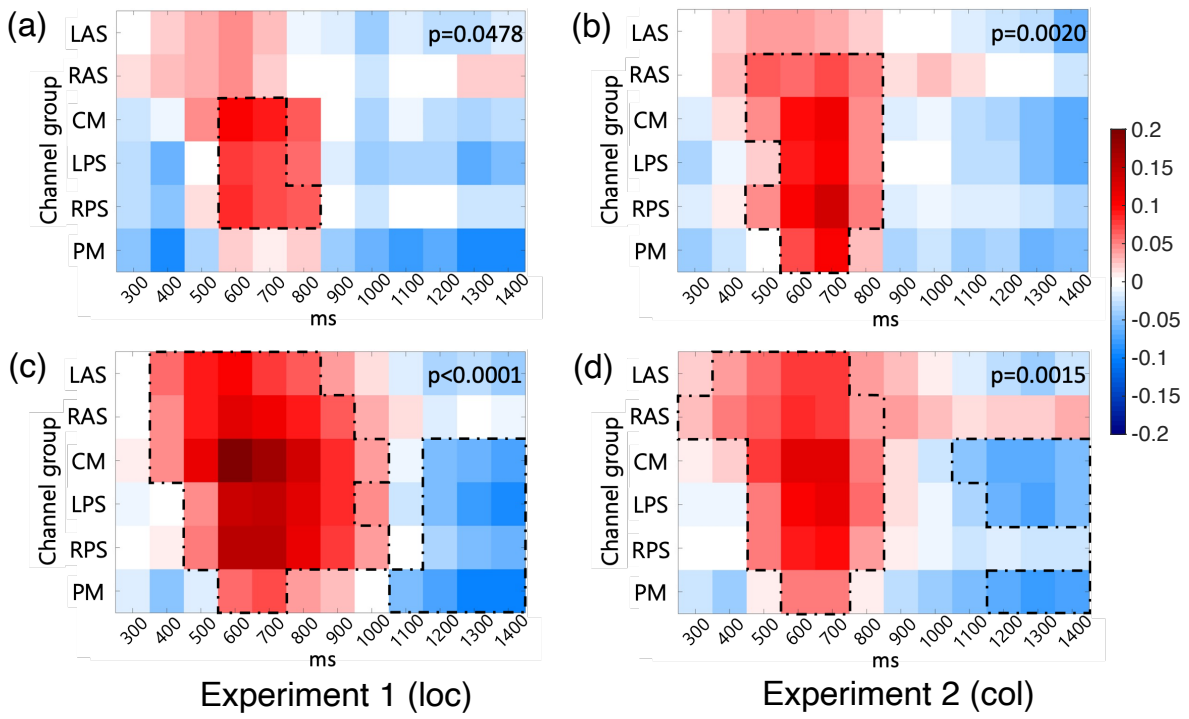


Figure 6.13. (a) The combined pattern of CR-SN vs. CR-MN, SC-RS vs. SI-RS with confidence control, and SI vs. CR with confidence control in Exp 1 and (b) in Exp 2 with mixing weights being the ratios of the coefficients in the CoSmIm model. (c) Mean difference between training classes of RS vs. F in Exp 1 (same as Figure 6.8(6-a)) and (d) Exp 2(same as Figure 6.8(6-b)).

and (d), the positive regions in the average patterns of the two figures are highly overlapped. In addition, the late negative part in Exp 2 in Figure 6.13(b) and the counterpart in Figure 6.13(d) are very similar.

6.3.2.4 Impact of Source Memory and Item Memory to Source Accuracy

Figure 6.14 illustrates the difference in AUROC removing either source memory or item memory for each subject and reveals how the source memory and item memory would change the source accuracy (defined as the number of SC-RS trials out of the number of SC-RS and SI-RS trials). The larger the difference between the AUROC of the CoSmIm model and the CoIm model, the more important the source memory is for the test subject. Likewise, The larger the difference between the AUROC of the CoSmIm model and the CoSm model, the more influential the item memory is for the test subject. The correlation of AUROC difference and source accuracy was

Table 6.11. The average coefficients and ratios of predictor variables in different linear regression models.

Model Parameter	Experiment 1			
	CoSmIm	CoSm	CoIm	SmIm
Conf. coef	0.6366	0.6374	0.6756	NaN
S.M. coef	0.5508	0.5487	NaN	0.5955
I.M. coef	0.0313	NaN	-1.579e-3	0.0466
Intercept		1.092e-4		
Conf. ratio	0.5223	0.5374	1.0023	NaN
S.M. ratio	0.4520	0.4626	NaN	0.9274
I.M. ratio	0.0257	NaN	-0.0023	0.0726
Model Parameter	Experiment 2			
	CoSmIm	CoSm	CoIm	SmIm
Conf. coef	0.6251	0.6221	0.6663	NaN
S.M. coef	0.1953	0.2005	NaN	0.3275
I.M. coef	0.3397	NaN	0.3425	0.3343
Intercept		4.283e-3		
Conf. ratio	0.5388	0.7563	0.6605	NaN
S.M. ratio	0.1683	0.2437	NaN	0.4949
I.M. ratio	0.2928	NaN	0.3395	0.5051

calculated using non-parametric Spearman correlation. In the location source experiment, higher source accuracy correlates with a bigger AUROC difference between the CoSmIm and CoIm models; while in the frame color source experiment, higher source accuracy only marginally correlated with a bigger difference between the CoSmIm and CoSm models.

6.3.2.5 Importance of Components to Different Behaviors

Figure 6.15 (d) shows the remember source (RS) false alarm rate is larger for subjects where the importance of the item memory component is lower. When it comes to the wrong source judgments where participants believed they remembered the source with correct item judgments, Figure 6.16 (a) and (d) show negative trends opposite of the trends in Figure 6.14 (a) and (d). That is for participants with a higher percentage of their SI trials labeled as RS (where they were incorrectly believing they remembered the source), the Source memory and Item memory components are less important at predicting RS vs F judgements in Exp 1 and 2,

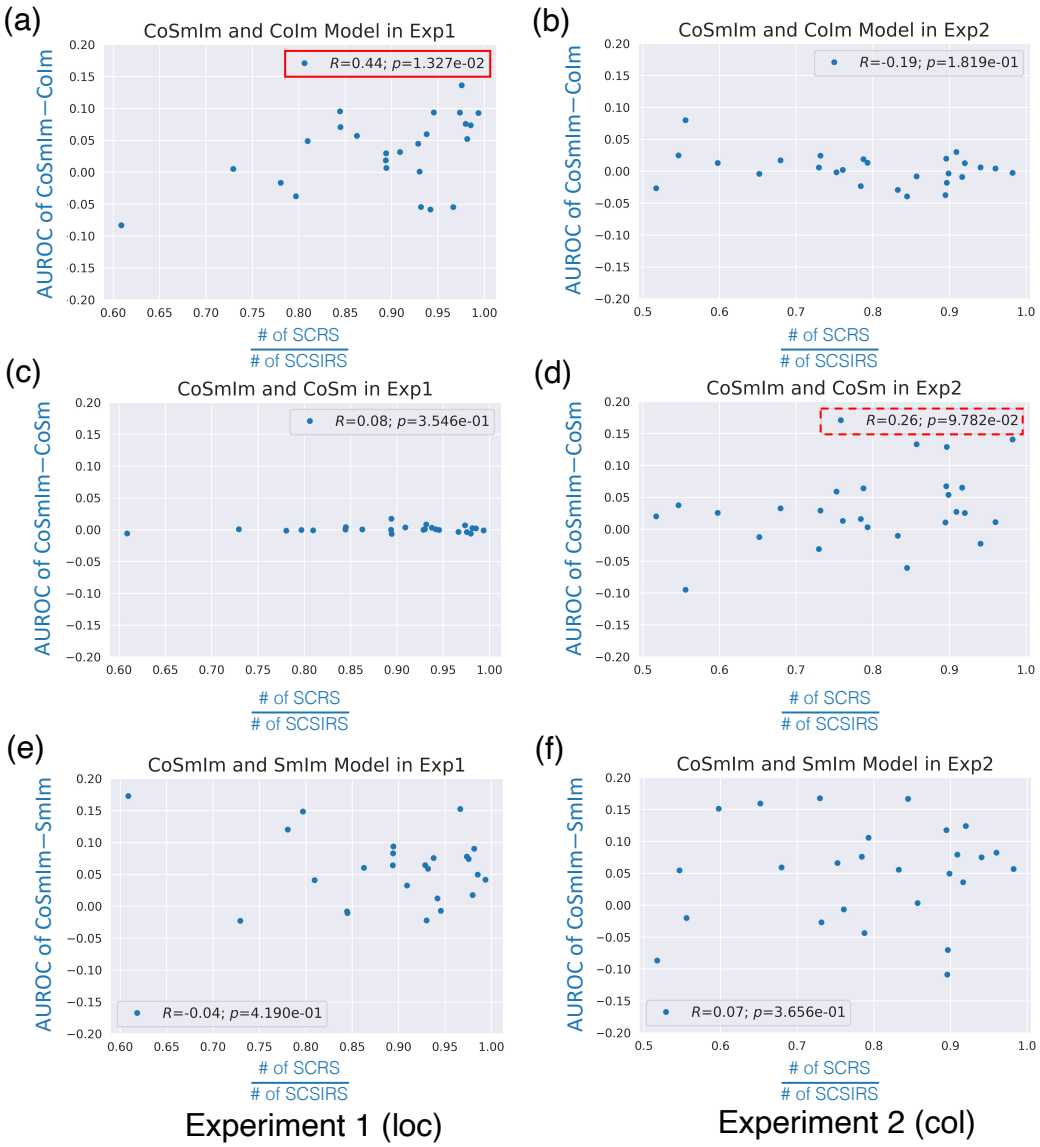


Figure 6.14. (a) AUROC of CoSmIm subtracted by AUROC of CoIm, showing the correlation between the effect of the source memory component to CoSmIm model and subjects' source accuracy in Exp 1 and (b) Exp 2. (c) AUROC of CoSmIm subtracted by AUROC of CoSm, showing the correlation between the effect of the item memory component to CoSmIm model and subjects' source accuracy in Exp 1 and (d) Exp 2. (e) AUROC of CoSmIm subtracted by AUROC of SmIm, showing the correlation between the effect of the confidence component to CoSmIm model and subjects' source accuracy in Exp 1 and (f) Exp 2.

respectively. Figure 6.17 (d) shows that the importance of item memory to fitting RS vs F EEG differences is positively correlated with the correct item recognition but wrong source judgment being labeled with remember other responses . In Figure 6.18 (d) and (f), the importance of

Table 6.12. The significance of the correlation between AUROC of CoSmIm model subtracted by AUROC of other models with two component and behavior ratio different from zero.

CoSmIm − Corr. with	Experiment 1			Experiment 2		
	CoIm	CoSm	SmIm	CoIm	CoSm	SmIm
$\frac{\text{SCRS}}{\text{SCSIRS}}$	+sig.	ns	ns	ns	ns	ns
$\frac{\text{FARS}}{\text{RS}}$	ns	ns	ns	ns	−sig.	ns
$\frac{\text{SIRS}}{\text{SI}}$	−sig.	ns	ns	ns	−sig.	ns
$\frac{\text{SIRO}}{\text{SI}}$	ns	ns	ns	ns	+sig.	ns
$\frac{\text{SCRO}}{\text{SCSIRO}}$	ns	ns	ns	ns	+sig.	−sig.

item memory is revealed to have a positive correlation with making remember other information responses when a subject correctly recognized the item and recollected the source information, whereas the importance of confidence is negatively correlated.

Table 6.12 summarizes the correlations between components and the behaviors selected in Figure 6.14 to 6.18. More discussion could be found in later sections.

6.3.3 Discussion

6.3.3.1 Similarity of LOSO Classifier and CoSmIm Model

The subject AUROCs of the CoSmIm model and the subject AUROCs of the RS vs. F LOSO classifier are highly correlated as shown in Figure 6.11, and the AUROCs for most subjects are comparable (blue dots are close to the diagonal line). This not only shows the CoSmIm model could achieve comparable performance in classifying RS and F but also the combination of the three components could provide a great explanation for the participant's difference between RS and F.

In the location source experiment, the projections of responses from the RS vs. F classifier in Figure 6.12(c) and the predicted projections from CoSmIm model in Figuer 6.12(a) show nearly identical relationships of the responses except for the difference in the scale. In the frame color source experiment, the projections from RS vs. F classifier in Figure 6.12(d) show even higher similarity in both the distribution and the scale to the projections from CoSmIm model in

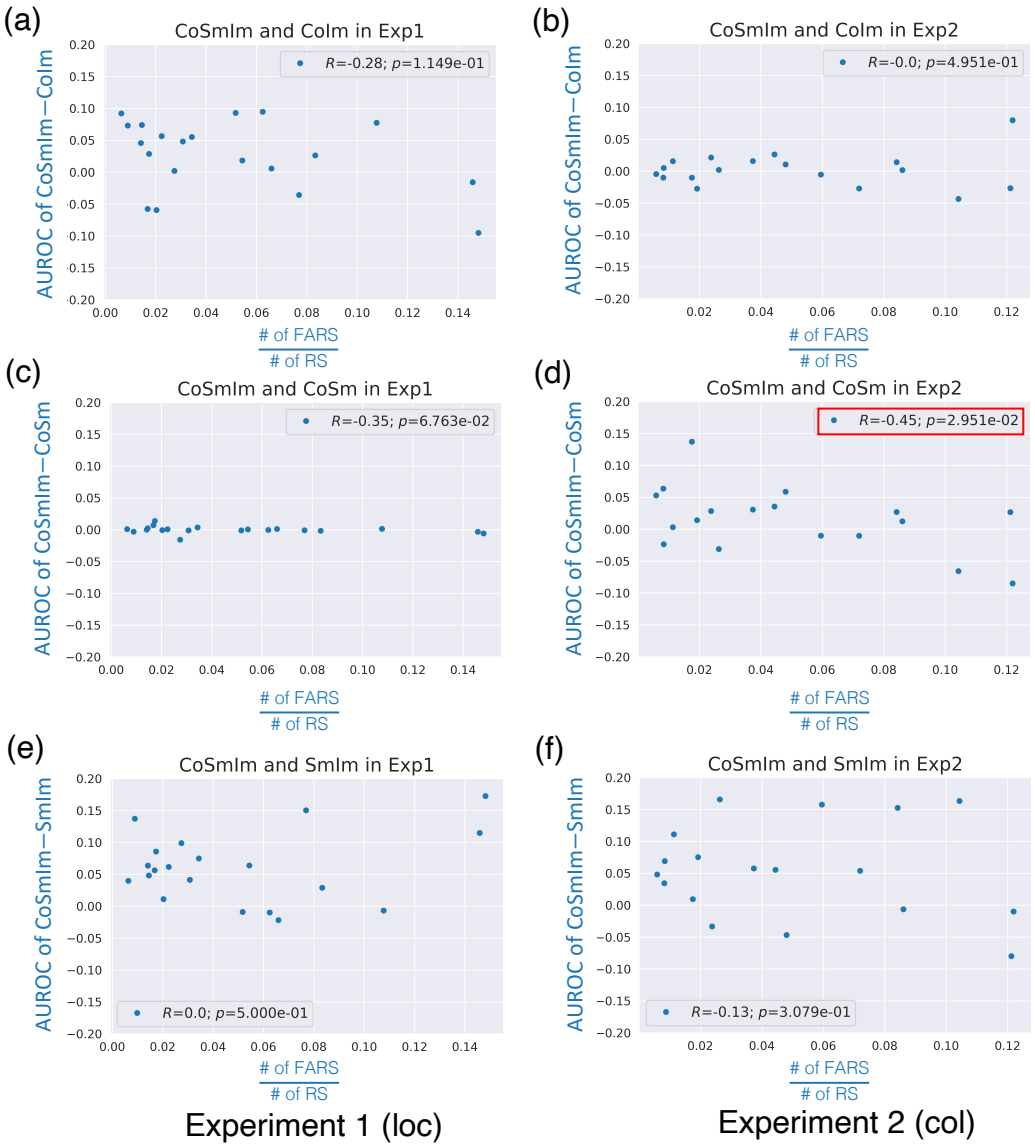


Figure 6.15. (a) AUROC of CoSmIm subtracted by AUROC of CoIm, showing the correlation between the effect of the source memory component to CoSmIm model and subjects' false item recognition with RS responses in Exp 1 and (b) Exp 2. (c) AUROC of CoSmIm subtracted by AUROC of CoSm, showing the correlation between the effect of the item memory component to CoSmIm model and subjects' false item recognition with RS responses in Exp 1 and (d) Exp 2. (e) AUROC of CoSmIm subtracted by AUROC of SmIm, showing the correlation between the effect of the confidence component to CoSmIm model and subjects' false item recognition with RS responses in Exp 1 and (f) Exp 2.

Figure 6.12(b).

For both source types, the patterns of RS vs. F in Figure 6.13 (c) and (d) show consistency

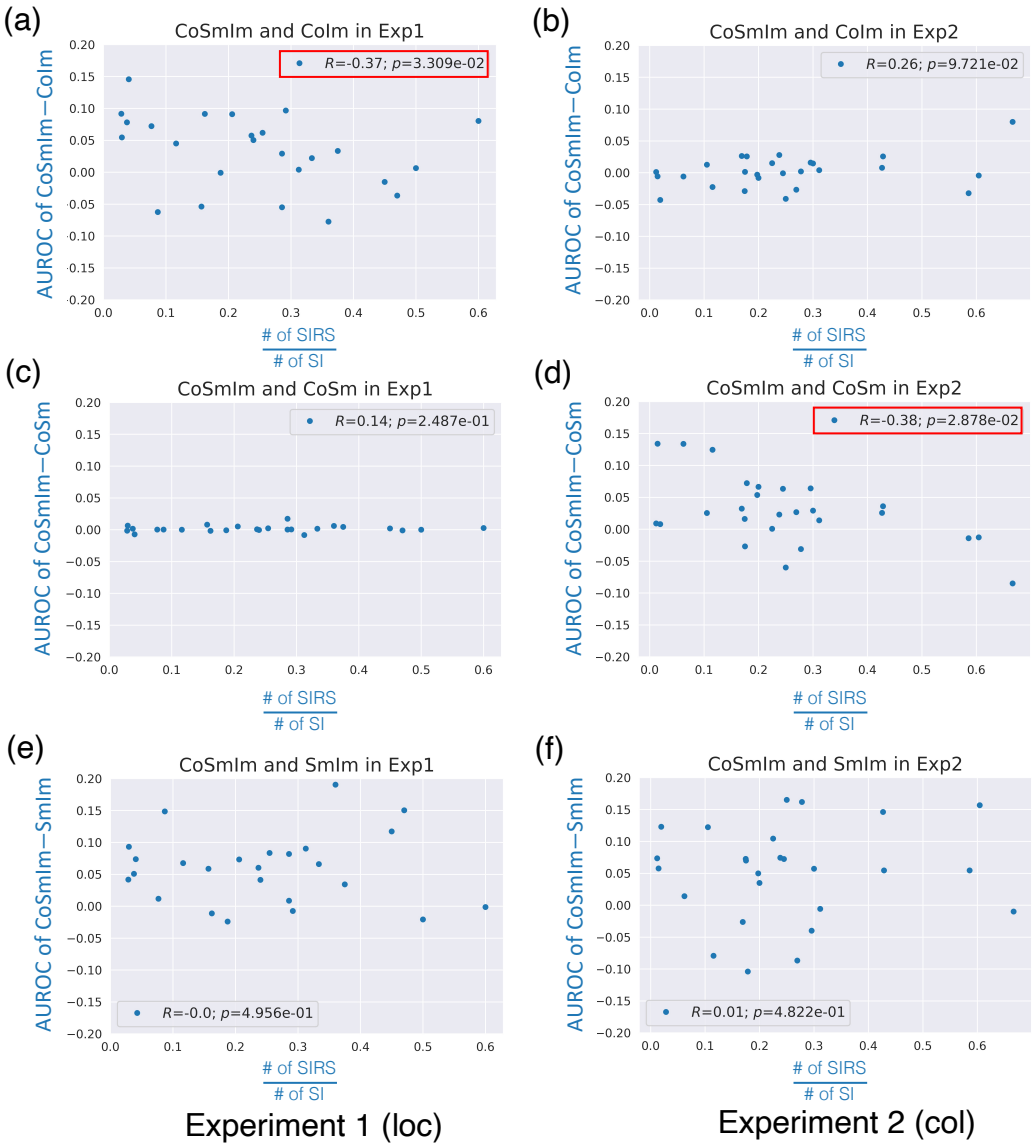


Figure 6.16. (a) AUROC of CoSmIm subtracted by AUROC of CoIm, showing the correlation between the effect of the source memory component to CoSmIm model and subjects' perceived correct source memory with SI responses in Exp 1 and (b) Exp 2. (c) AUROC of CoSmIm subtracted by AUROC of CoSm, showing the correlation between the effect of the item memory component to CoSmIm model and subjects' perceived correct source memory with SI responses in Exp 1 and (d) Exp 2. (e) AUROC of CoSmIm subtracted by AUROC of SmIm, showing the correlation between the effect of the confidence component to CoSmIm model and subjects' perceived correct source memory with SI responses in Exp 1 and (f) Exp 2.

with the patterns of CoSmIm models in Figure 6.13 (a) and (b), respectively. For the location source experiment, the positive pattern in CoSmIm pattern is highly overlapped with the positive

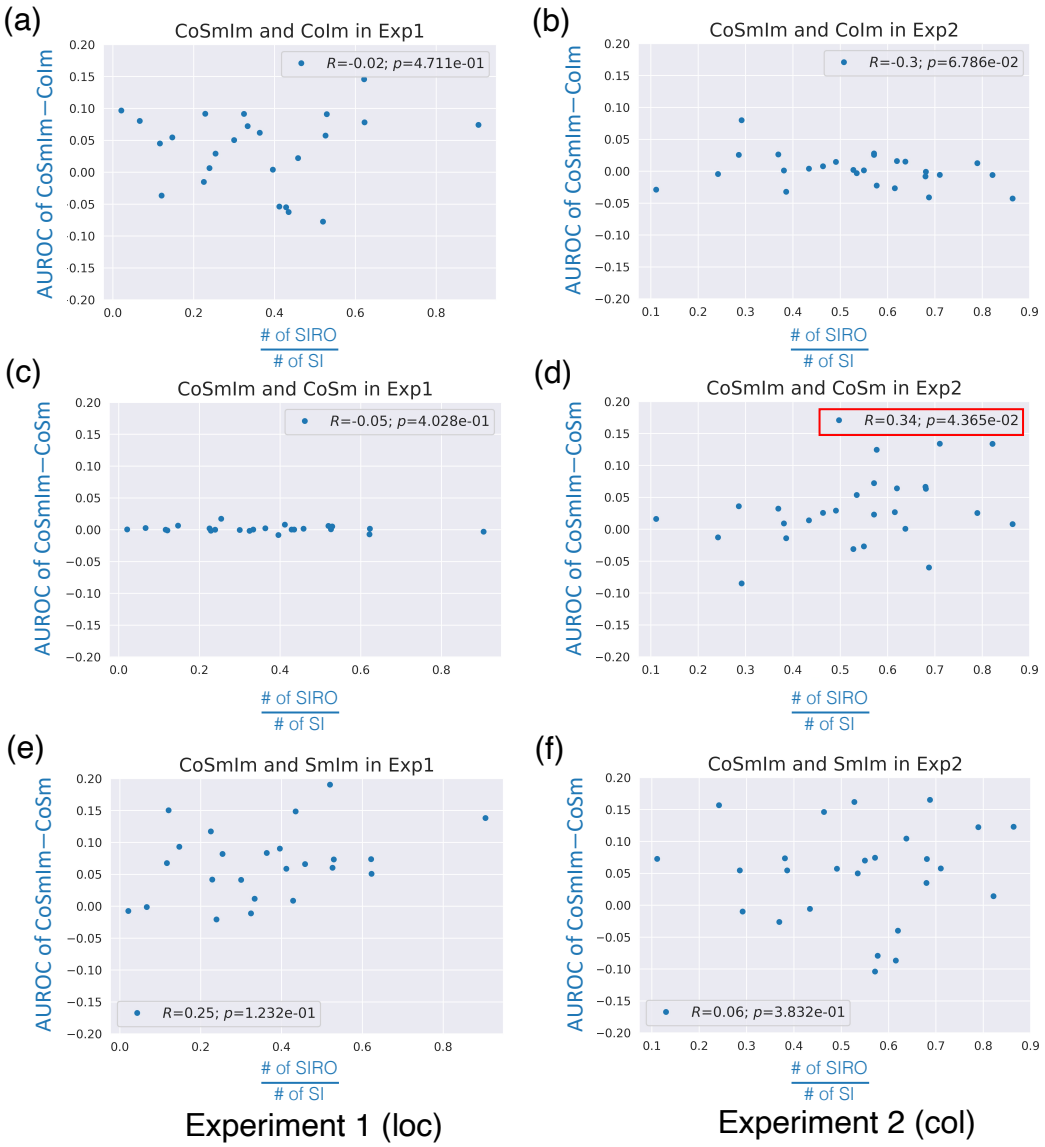


Figure 6.17. (a) AUROC of CoSmIm subtracted by AUROC of CoIm, showing the correlation between the effect of the source memory component to CoSmIm model and subjects' recollected other source responses with wrong studied source judgments in Exp 1 and (b) Exp 2. (c) AUROC of CoSmIm subtracted by AUROC of CoSm, showing the correlation between the effect of the item memory component to CoSmIm model and subjects' recollected other source responses with wrong studied source judgments in Exp 1 and (d) Exp 2. (e) AUROC of CoSmIm subtracted by AUROC of SmIm, showing the correlation between the effect of the confidence component to CoSmIm model and subjects' recollected other source responses with wrong studied source judgments in Exp 1 and (f) Exp 2.

part in RS vs. F pattern. The only incongruent region is the late negative block over parietal electrodes. In fact, the patterns of the three components (confidence, source memory and item

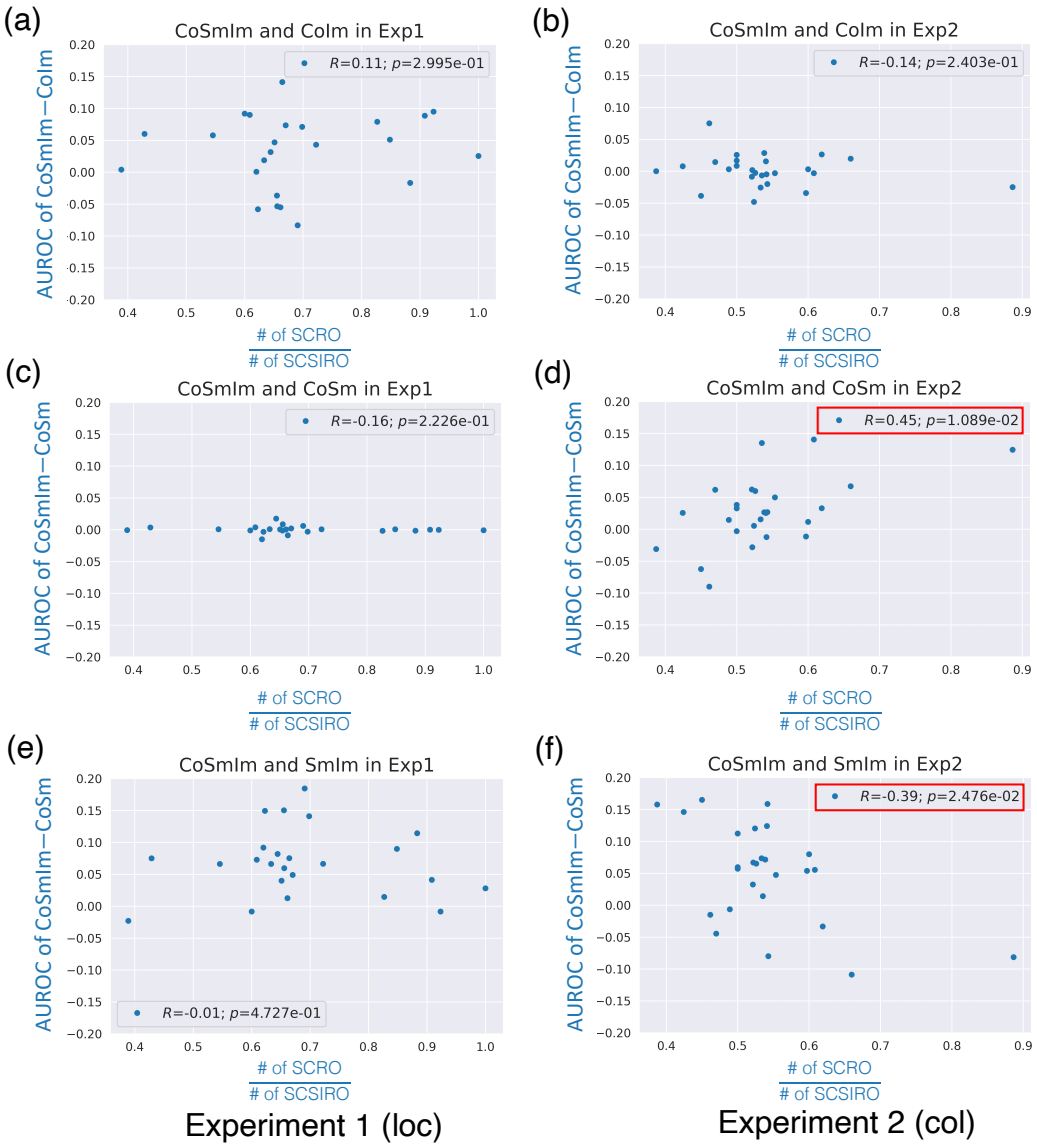


Figure 6.18. (a) AUROC of CoSmIm subtracted by AUROC of CoIm, showing the correlation between the effect of the source memory component to CoSmIm model and subjects' source accuracy with RO responses in Exp 1 and (b) Exp 2. (c) AUROC of CoSmIm subtracted by AUROC of CoSm, showing the correlation between the effect of the item memory component to CoSmIm model and subjects' source accuracy with RO responses in Exp 1 and (d) Exp 2. (e) AUROC of CoSmIm subtracted by AUROC of SmIm, showing the correlation between the effect of the confidence component to CoSmIm model and subjects' source accuracy with RO responses in Exp 1 and (f) Exp 2.

memory with confidence control) do not include a clear cluster for this block. There could be additional cognitive processes needed to consider to reconstruct the late negative block or could

be more variability between subjects. As for the frame color source experiment, the CoSmIm pattern and the RS vs. F pattern are more alike except for the sizes of the significant clusters.

The similarity between patterns shows that the utilized spatio-temporal features for the classification in RS vs. F LOSO classifier and the CoSmIm model are consistent, and the projections of the non-trained responses being almost identical using the RS vs. F LOSO classifier and the CoSmIm model further justify that.

6.3.3.2 Importance of Source Memory and Item Memory in CoSmIm model

From the coefficients of the CoSmIm models in Table 6.11, the confidence component is the most important factor contributing to the difference between remember and familiar responses.

For the location source experiment, the coefficients for the confidence and the source memory components are higher than the one for item memory component, indicating that both confidence and source memory are more important in making a remember decision in Exp 1. The same conclusion could also be drawn from Table 6.10 that the performance drops significantly when using CoIm model instead of CoSmIm model.

For the frame color source experiment, while the confidence component still gets the highest weight, the item memory component now gets higher weight than the source memory component, implying that item memory is more important than source memory when making the remember responses in Exp 2. Accordingly, the drop of AUROC from using CoSmIm to CoSm model in Table 6.10 also indicates that item memory is indispensable in making remember judgments in the color frame source experiment. In the color frame source experiment, the performance of correct source retrieval is not as good as in the location source experiment (see Table 6.8), and the R-K judgments are revealed to be more dependent on item memory. Since the source memory is poor on this task, it is possible that confident recognition of the studied item leads the subjects to believe that they actually remember the color of the frame and make the R responses.

In Exp 1, the strong correlation of the subject AUROC difference (between CoSmIm and CoIm models) and the subject source accuracy (defined as the number of SCRS trials over the number of SCRS and SIRS trials) where R is significantly different from 0 in Figure 6.14(a) suggests that there is a strong correlation between the importance of the source memory in accounting for a subject's remember vs familiar responses and their degree of source accuracy. In Exp 2, although item memory is more important than source memory to the CoSmIm model, the correlation of the importance of item memory and the source accuracy in Figure 6.14(d) is positive but only marginally significantly different from 0. This could come from the increased difficulty and lower overall source accuracy in the frame color source experiment.

6.3.3.3 Ablation Study for Difference in Behaviors

For the color frame source experiment, the frequency of the false alarm with remember source responses increases as the difference between AUROCs from the CoSmIm model and from the CoSm model decreases as shown in Figure 6.15. While the difference between the AUROCs could be interpreted as the importance of the item memory component to the model, it could also be viewed as how reliable item memory is for a subject when making remember decisions. As a result, the negative correlation between false alarm rate in remember source responses and the difference in AUROCs indicates that subjects believed they remembered the source could actually be wrong if they made the judgment not based on their item memory.

In the previous section Section 6.3.3.2, the correlation between source accuracy and source memory for Exp 1 and and the correlation between source accuracy and item memory for Exp 2 were discussed. For SI (source incorrect) responses, the item recognition was correct but the source judgment was wrong. In Figure 6.16 (a) and (d), source memory and item memory suppresses subjects' tendency to make remember source responses when they actually don't remember the source in Exp 1 and 2, respectively. The suppression could also be thought as the awareness of wrong source judgment, and source memory and item memory contribute it in a similar way how they affects source accuracy in Exp 1 and 2, respectively.

For Exp 2, the proportion of SI-RO to SI is positively correlated with the item memory as shown in Figure 6.17 (d). Following the idea that confident item recognition could lead to remember source responses as mentioned in Section 6.3.3.2, the positive correlation could also be explained as the item recognition leads to recollection responses. Addition to that, making remember other information responses might also reflect the awareness of wrong source judgments.

The proportion of SC-RO to SC&SI-RO could be interpreted as an index for being conservative of the subjects' source judgment because they actually remember the source. For the color frame source experiment, the index is revealed to be positively correlated with item memory but negatively correlated with confidence as illustrated in Figure 6.18 (d) and (f). While item memory has been shown to associated with source accuracy for RS responses, the observation still holds for RO responses. The negative correlation between the index and the confidence in Exp 2 suggests that the more conservative of a subject about their source judgment, the less they rely on the confidence component.

6.4 General Discussion

6.4.1 Subject-Dependent vs. Subject-Independent Training

Most of the memory studies using single-trial EEG classification were based on subject-dependent classifiers trained for each individual subject (Noh et al., 2014; Ratcliff et al., 2016; Noh et al., 2018; Kim, Jeong, Kim, & Chung, 2020) and demonstrate the efficacy of using subject-dependent classifiers for EEG analysis. Therefore, subject-dependent is often considered the benchmark for single-trial EEG studies in memory. However, subject-dependent classifiers could be trained to fit idiosyncratic features and not able to generalize to common features across subjects. Moreover, for behaviors without enough trials in a subject (i.e. SI-RS in this study), the feasibility of using subject-dependent classifier is restricted.

On the other hand, subject-independent classifiers (Liao et al., 2018) showed comparable

performance to within subject classifiers by combining training data across subjects. In this study, the LOSO classifier even achieves better performance than the LOTO classifier (see Table 6.9 and Figure 6.11). In addition, behaviors having a limited number of trials could now be explored due to having sufficient training data pooled across different subjects.

6.4.2 The Components for Confidence in Correct Old and New Responses Are the Same

Previous ERP studies (Addante et al., 2012; Woodruff, Hayama, & Rugg, 2006; Wynn et al., 2020; Sarah & Rugg, 2010) have shown that the LPC is correlated with high confidence recognition responses. In this chapter, the mean difference pattern of the CR-SN vs. CR-MN (see Figure 6.7 (1)) contains a positive cluster located over parietal areas around 600-800 ms which overlaps with the spatio-temporal features of the LPC effect. The same LPC effect could also be observed in the patterns of RS vs. F in Figure 6.8 (6), but the effect is relatively weak for SC-RS vs. SI-RS with confidence controlled in Figure 6.7 (3) and discrepant for SI vs. CR with confidence controlled in Figure 6.8 (5). In addition, we show that the confidence classifier trained on correct new responses could distinguish between remember and familiar responses. Furthermore, we also show that using linear regression models, the confidence difference between remember and know responses could also be represented by the confidence difference in the correct new responses. These results support the findings in Schwarze, Bingel, Badre, and Sommer (2013); Rutishauser et al. (2018) that the confidence in correct old responses and the confidence in correct new responses could reflect overlapping components.

6.4.3 Linear Regression Models for Decomposing Cognitive Components

In this chapter, we hypothesize that the difference between remember and familiar responses is attributed to the difference in confidence, source memory, and item memory. We first trained the CR-SN vs. CR-MN confidence classifier, SC-RS vs. SI-RS source memory classifier with confidence control, and SI vs. CR item memory classifier with confidence control

and showed that they could all separate RS from F. We then trained the RS vs. F classifier and the linear regression model to approximate the projections onto the RS vs. F classifier with the projections from the three original classifiers. Note that the way we adopted confidence control for training the source memory and item memory classifier could resolve the issue of co-linearity between predictors, and the weights in the trained linear regression model (e.g. confidence, source memory, item memory) could then reflect the importance of each term to the prediction (e.g. R-K). This method is not limited to the model for interpreting the difference between remember and know responses. In fact, it could be applied to study any cognitive processes that could possibly be decomposed into multiple components. Our analysis was restricted to time domain signals from specific channel groups know to be involved in frontal, parietal, and late posterior old/new effects. It is possible that using the frequency domain information from multiple electrode location and frequency bands could improve the performance of the classifiers. However, the variation of spectrum in memory across subjects (Klimesch, 1997) would required extensive study in order to extract reliable features for training the classifiers.

6.5 Revisiting Source Memory in Experiment 2 and 3

Earlier in this chapter, we showed that the CoSmIm models fitted differently for different source information because the color boundary information was more difficult to remember than the location source information. However, comparing to Exp 1, the lesser importance of the source memory component in the CoSmIm model for Exp 2 could have stemmed from the worse source recollection performance and the consequently worse source memory classifier. In order to justify this assumption, we are going to use the source memory, item memory, and confidence classifiers trained on Exp 1 as the three components and learn a CoSmIm model based on these three components from Exp 1 to fit the R-K difference in Exp 2.

In addition, we would like to know if we could obtain similar CoSmIm models for experiments using the same type of source information. To this end, we will apply the same

method to learn CoSmIm models with the same three components found in Exp 1 to fit the R-K differences in Exp 3-location and 3-color.

6.5.1 Methods

In the analyses for this section, the source memory, item memory, and confidence classifiers were trained on Exp 1 to represent the three components, but the R-K classifier was trained on each experiment (e.g., Exp 2, Exp 3-loc, and Exp 3-col) where the R-K difference was supposed to be learned by the CoSmIm model. For Exp 2, the SC-RS and SC&SI-F trials were projected onto the three classifiers trained on Exp 1, and these projections were then treated as the independent variables and represented the three components in the CoSmIm model. The projections of the same trials from the RK classifier trained on Exp 2 were the dependent variable that the CoSmIm model was trained to predict. The identical training procedures were adopted to train separate CoSmIm models for Exp 3-loc and Exp 3-col.

6.5.2 Results

Table 6.13 shows the AUROCs of R-K LOSO classifiers and the CoSmIm linear regression models with components from Experiment 1 to predict R-K judgements in Experiment 2, 3-location, and 3-color. In Exp 2, the AUROC of the CoSmIm model using components from Exp 1 is really close to the AUROC of the CoSmIm model using components from Exp 2 (see Table 6.9) and significantly better than the LOSO classifier. This suggests that either using the components from Exp 1 or Exp 2, the CoSmIm model could interpret the R-K difference in terms of source memory, item memory, confidence components really well. For Exp 3-loc and 3-col, the AUROCs of RS vs. F from the CoSmIm models with components from Exp 1 are comparable to the AUROCs of the LOSO classifiers but lower, similar to the results for Exp 1 (see Table 6.9).

The comparisons between the 3-component model with components from Exp 1 and 2-component models with components from Exp 1 based on the AUROCs of RS vs. F are shown

Table 6.13. The average AUROCs of classifying SC-RS and SC&SI-F using different classifiers and linear regression models with components from Exp 1 (shown as Linreg (1)).

AUROCs of RS vs. F	LOSO Classifier	CoSmIm Linreg (1)	LOTO Classifier	CoSm Linreg (1)	CoIm Linreg (1)	SmIm Linreg (1)
Exp 2	0.6315^{**}	0.6547^{**}	0.6137^{**}	0.6437^{**}	0.6276^{**}	0.6052^{**}
Exp 3-loc	0.6601^{**}	0.6272^{**}	0.6536^{**}	0.6265^{**}	0.5953^{**}	0.5922^{**}
Exp 3-col	0.6330^{**}	0.6104^{**}	0.6404^{**}	0.6054^{**}	0.5814^{**}	0.5768^{**}

** $p < 0.01$,* $p < 0.05$, † $p < 0.1$

Table 6.14. The p-values of the paired one-tailed Student's t-test between subject AUROCs of different linear regression models with components from Exp 1.

p-value of AUROCs: CoSmIm (1)	vs. CoSm (1)	vs. CoIm (1)	vs. SmIm (1)
Experiment 2	9.736e-3	4.605e-3	0.0135
Experiment 3-loc	0.7107	1.344e-4	1.585e-3
Experiment 3-col	0.1996	6.103e-5	0.0287

in Table 6.14. For Exp 2, the AUROC would significantly drop if any of the components from Exp 1 is removed from the CoSmIm model. This indicates when using components from Exp 1 to learn the CoSmIm linear regression model, all the components are important to interpret the R-K difference in Exp 2. For Exp 3-loc and Exp 3-col, the AUROCs only drop significantly when removing the source memory component or the confidence component but not the item memory component. This is in accordance with the findings in Exp 1 in Section 6.3.2.1, where removing the item memory component does not significantly change the performance of the linear regression model.

The coefficients for components from Exp 1 in the CoSmIm models in Exp 2, 3-loc, and 3-col are shown in Table 6.15. The coefficient for the confidence component are consistently higher than the coefficients for source memory and item memory components in the CoSmIm model for each experiment, which indicates the importance of the confidence component in the CoSmIm model. For the CoSmIm model with components from Exp 1 in Exp 2, the coefficient for the source memory component is higher than the coefficient for the item memory component; this is different from the earlier CoSmIm model for Exp 2 in Table 6.11 where the coefficient for

Table 6.15. The average coefficients and ratios of predictor variables in different linear regression models using components from Exp 1 (shown as CoSmIm (1)).

CoSmIm (1) parameters	Experiment 2	Experiment 3-loc	Experiment 3-col
Conf. coef	0.6384	0.2846	0.2704
S.M. coef	0.2618	0.2366	0.1618
I.M. coef	0.1366	0.0380	0.0875
Intercept	-2.214e-4	2.029e-3	2.038e-4
Conf. ratio	0.6143	0.5890	0.5191
S.M. ratio	0.2542	0.4897	0.3105
I.M. ratio	0.1315	-0.0787	0.1704

the item memory component is higher than the counterpart for the source memory component.

6.5.3 Discussion

In Section 6.3.3.2, we claimed that the item memory component was more important than the source memory component in the CoSmIm model in Experiment 2. We assumed the source memory component and the item memory component from Exp 2 were both reliable because the AUROCs of the source memory classifier and the item memory in Table 6.5 were both significantly above 0.5 with p-values smaller than 0.05 and 0.01 respectively. However, when using the components from Exp 1 to learn the CoSmIm for Exp 2, the source memory component instead becomes more important than the item memory component. It is possible the source memory classifier in Exp 2 does not perform well enough to reflect a clear source memory component because, even though the AUROC is significantly above 0.5, it is still close to 0.5 compared to the counterpart in Exp 1. Another possible explanation for this discrepancy between using components from Exp 1 and Exp 2 could be the item memory classifier is better trained for Exp 2 than Exp 1. However, it is less convincing than the previous explanation because the AUROC of the item memory classifier in Exp 1 is higher than the AUROC of the source memory classifier in Exp 2, which means the item memory component in Exp 1 could be more well-defined than the source memory component in Exp 2.

For both Experiment 3-location and 3-color, the CoSmIm models with components from

Exp 1 rely more on the source memory component according to the coefficients in Table 6.15, which is consistent with the null effect when comparing the AUROCs of CoSmIm and CoSm in Table 6.14. This result is also consistent with the finding in Exp 1 in Section 6.3.3.2. The similarity in the lack of item memory component in R-K difference in Exp 1 and Exp 3-col is interesting because item memory component was crucial in R-K difference in Exp 2 (color source) but not in Exp 1 (location source), and the source accuracy in Exp 3-col, even with only two color implemented, is not comparable to the counterpart in Exp 1 (see Table 6.16). One possible explanation is that in Experiment 2, the subjects adjust their procedure for deciding "remember color" due to the difficulty of the task. (Failing to remember the color, they associate strong item memory with a belief in remembering the source information). The behavioral results show that they are rarely able to remember the color outline. In Experiment 3, however, the same subjects are completing both the color and location task and so they are more likely to keep a consistent procedure that more accurately assesses their memory of the source. That is during the interleaved location source sessions and color source sessions, the subject learned how to recognize their use of the source information component in R-K judgments during location source sessions and applied the same strategy to color source sessions.

Table 6.16. Total numbers of trials of correct item responses in Exp 3-loc and 3-col.

# of trials	SC-RS	SI-RS	SC-RO	SI-RO	SC-F	SI-F	CR-MN	CR-SN
Exp 3-loc	2720	364	907	464	726	577	947	1409
Exp 3-col	2030	593	1068	864	726	680	921	1500

	SC/SC&SI	SC-RS/SC&SI-RS
Exp 3-loc	0.7560	0.8820
Exp 3-col	0.4938	0.7739

6.6 Chapter Acknowledgements

Chapter 6, in full, is the reorganized version of the manuscript in preparation. Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. "Using single-trial EEG classifiers to decompose remember-know difference", *in preparation*. The dissertation author is the primary

author of this manuscript.

Chapter 7

Memory Model Consists of Source Memory, Item memory, and Confidence

In the previous chapter (Chapter 6), we showed that the difference between remember and know responses could be decomposed by three components: source memory, item memory, and confidence. These three components do not entirely match the most frequently used single-process or dual-process models. In this chapter, we would like to review some single-process and dual-process models, show why the confidence component cannot be explained by either of these two models, and propose our own model.

7.1 Remember-Know (RK) Paradigm

In our study, a modified Remember-Know paradigm was employed to obtain a subjective rating after the subject made an old judgment to the first question. To discuss the following memory models, a short revisit of the RK paradigm would be beneficial.

The remember-know paradigm was first introduced by Tulving (1985) to measure the states of awareness during memory retrieval; remember (or R) responses were associated with retrieval from episodic memory while know (or K) responses were associated with retrieval from semantic memory. Gardiner (1988) later refined the RK paradigm: Remember responses were taken to reflect awareness of some aspect of what had occurred or had been experienced the time the test item was first presented (during initial encoding); Know responses were taken to

reflect recognition of the test item without the ability to recollect (without awareness of) what had happened or had been experienced when the test item was first presented.

7.2 Single-Process and Dual-Process Model

Some researchers (W. Donaldson, 1996; Hirshman & Henzler, 1998) have proposed that R and K responses reflect different levels of confidence or memory strength during memory retrieval; this view is called the *single-process interpretation* because the same process is assumed to explain all responses, old and new. On this view during recognition, when memory strength of a test item exceeds a more stringent criterion, an R response is made; if the memory strength falls between a stringent criterion and a loose criterion, a K response is made; and if the memory strength is less than the loose criterion, a "new" response is given. Memory strength is assumed to reflect the confidence level for judging that an item is old; in short, R responses correspond with high to very high levels of confidence, and K responses correspond to a low level of confidence. A common way to visualize the single-process model and its various behavioral thresholds for recognition is via signal-detection theory as shown for the Remember-Know paradigm in Figure 7.1 (a) (W. Donaldson, 1996).

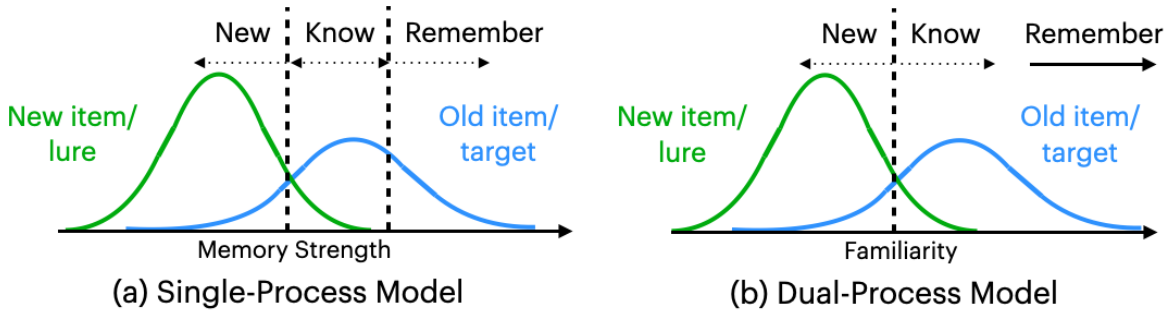


Figure 7.1. (a) Single-process and (b) dual-process models for Remember-Know paradigm.

In contrast to the single-process interpretation, more frequently studies using R-K paradigm assume that R and K responses reflect different types of memory retrieval, based on different processes. This view, consistent with Tulving's original proposal, is known as a

dual-process interpretation. While Tulving's proposal argues that R responses reflect episodic memory whereas K responses reflect semantic memory, a more recent popular proposal by (Jacoby, Yonelinas, & Jennings, 1997; Yonelinas et al., 2002) suggests that R responses are associated with a *recollection* process whereas K responses are associated with a *familiarity* process. In the recollection and familiarity based dual-process model, recollection and familiarity are two distinct processes where recollection is an all-or none component, and familiarity is a continuous variable; this dual-process model can be explained by signal-detection theory as shown in Figure 7.1 (b).

7.3 Problems of Single- and Dual-Process Models

In Chapter 6, we used CR-SN, CR-MN, SC&SI-F, SI-RS, and SC-RS to train classifiers to represent confidence, source memory, and item memory components as well as the difference between R and K responses. Applying the single-process interpretation and dual-process interpretation for these selected behaviors, we could present their distribution as in Figure 7.2 (a) and (b), respectively.

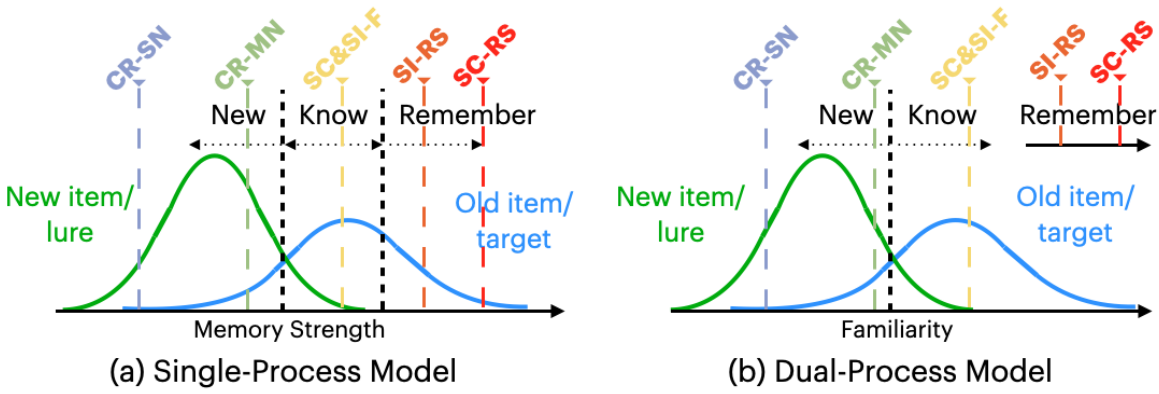


Figure 7.2. Behaviors selected for training the classifiers in Chapter 6 presented in the view of (a) Single-process and (b) dual-process models for Remember-Know paradigm.

According to single-process models, confidence represents differences in familiarity or memory strength between new (e.g., CR in our study), old-familiar (i.e., old items judged as familiar; e.g., SC&SI-F), and old-recollected (i.e., old items judged as recollected; e.g., SC-RS)

items. For new items, low familiarity is associated with high confidence and high familiarity is associated with low confidence. For old items, low familiarity is associated with low confidence and high familiarity is associated with high confidence. In other words, familiarity and confidence are positively correlated for old items, but negatively correlated for new items. We show that a classifier trained to differentiate high and low confidence new items (CR-SN vs. CR-MN classifier) can also differentiate high and low confidence old items (SC-RS and SC&SI-F), but the underlying signal is positively correlated with confidence for both old and new items, unlike the single-process prediction. Thus, the underlying confidence signal is unlikely to reflect item familiarity. Furthermore, it is unlikely to reflect recollection, because recollection should be absent for new items.

According to dual-process models, familiarity explains confidence differences between new (CR) and old-familiar (SC&SI-F) items, as explained above. However, confidence differences between old-familiar (SC&SI-F) and old-recollected (SC-RS) items are due to qualitative differences between familiarity and recollection processes. Critically, a single process should not be able to account for both differences between low and high confidence new items as well as between familiar and recollected items. However, as explained above, we found a confidence signal that was able to account for both low/high confidence new items and familiar/recollected old items, in contrast to the dual process perspective.

7.4 Our Proposed Model

Overall, our results are inconsistent with both single-process and dual-process views of confidence, which hold that confidence arises from the underlying familiarity and/or recollection processes. However, confidence is not enough to explain the difference between old-recollected (SC-RS) and old-familiar (SC&SI-F) items because the AUROC of Remember vs. Know responses from our confidence classifier is much lower than the counterpart from the Remember vs. Know classifier, which suggests there could be more components that could enhance the

separation of old-recollected (SC-RS) and old-familiar (SC&SI-F) items. In Chapter 6, we showed that adding source memory and item memory components together with a confidence component could explain the difference between Remember and Know responses very well.

In order to model our findings and remedy the inadequacies of the extant models, we propose a new model comprising the three major components investigated in Chapter 6: source memory, item memory, and confidence. Figure 7.3 shows this three-component model. With this model, we can explain the difference between behavior-pairs in each of the three components or their combination. The decision confidence (CR-SN and CR-MN difference), which is here considered as a component that is similar for old and new decisions, could be used to explain the confidence difference between old-recollected (SC-RS) and old-familiar (SC&SI-F) items. In addition, decision confidence can explain the confidence bias observed in the familiarity classifier in Chapter 5.4.2. The source memory component could be associated with the recollection process, except it is separate from the decision confidence component. As for the item memory component, it could be associated with the familiarity process but separate from the confidence component.

This three-component model is proposed based on our observation and findings in the current study. In the location source experiment, we found source memory and confidence were the two major components that accounted for the Remember-Know difference, whereas in the color source experiment, item memory and confidence were revealed to be the significant components to explain the difference between Remember and Know responses. As a result, all three components are indispensable in order to interpret Remember-Know differences for various source types.

Based on our model, during memory retrieval, subjects will use item memory to identify the studied items and the lure items. For studied items, low source memory, item memory, and confidence will lead the subjects to make familiar responses, whereas high source memory, item memory, and confidence will lead to remember responses. For the remember responses, source memory and/or confidence contributes to the correctness of the source information

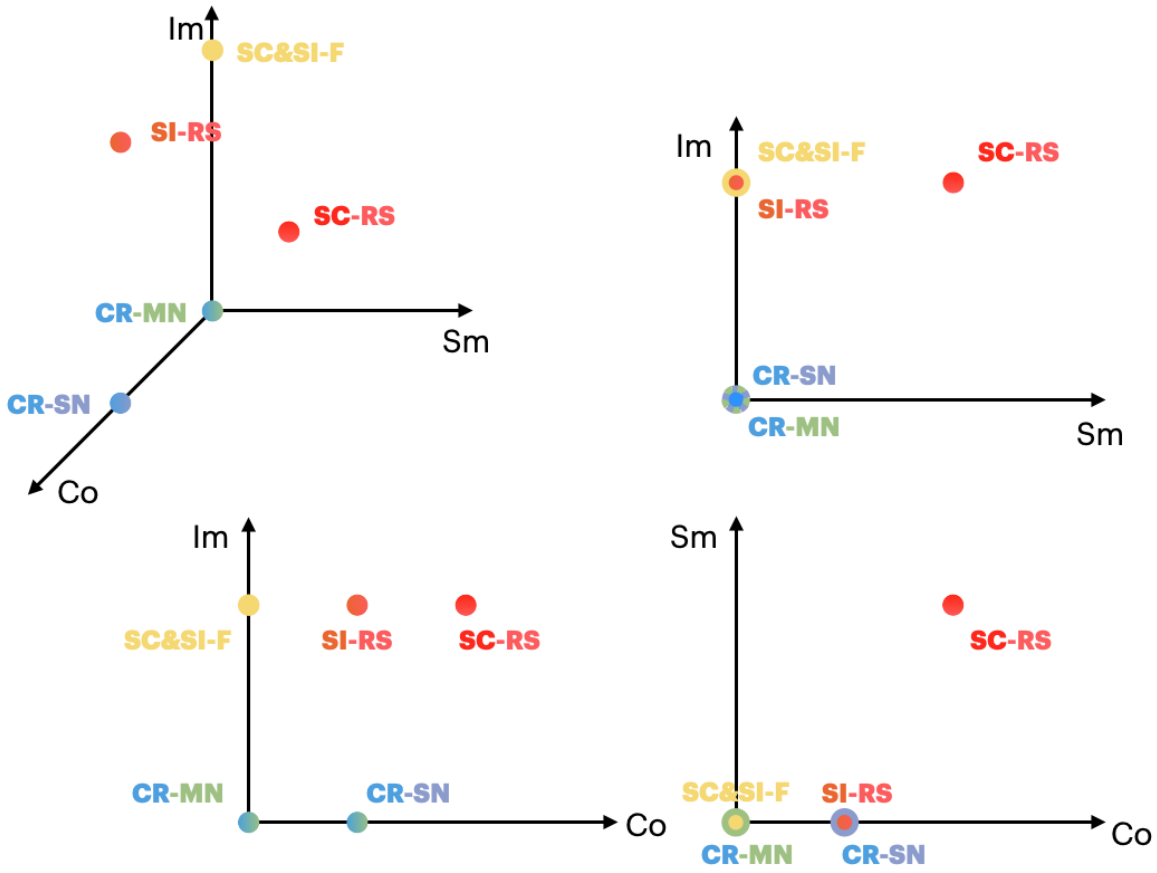


Figure 7.3. Illustration of our proposed model to explain memory behaviors in three components: source memory, item memory, and confidence. Locations of the behaviors are based on Exp 1.

judgement (difference between SC-RS and SI-RS) depending on the source type (e.g., confidence appears in SC-RS vs. SI-RS in Exp 1-loc but not 2-col). In Exp 1-loc, subjects who have better source memory performance (percentage of old-recollected (SC&SI-RS) judgements that are correct(SC-RS) - source hits) have more dependence on source memory for explaining the R-K EEG difference. In addition subjects who had low over-perceived recollection (percentage of incorrect source judgements (SI) they believed they had correct (SI-RS) source) also had more dependence on source memory for explaining the R-K EEG difference. These two correlations imply that in Experiment 1, the source memory response is related to behavioral ability to correctly remember the source. In Experiment 2, the item memory response plays a similar role. For correctly identified lure items, subjects will use the same confidence component (as used for

the old items) to determine their confidence levels of the new judgments.

The combination of the source memory component, the item memory component, and the confidence component could consistently explain R-K EEG difference across different experiments as shown in Section 6.5, where we successfully used components obtained from the location source experiment (Exp 1) to interpret the R-K EEG difference in color outline source experiment (Exp 2). In fact, the source memory component could be better learned in Experiment 1 where the source information was easier to remember and where subjects had higher source accuracy (Exp 1). By using the source memory component from Exp 1, we showed that the source memory component is actually indispensable in interpreting the R-K EEG difference across different source types, which we failed to show when using the (less-well learned) source memory component from the color source experiment (Exp 2). In addition, the item memory component could have been less important in explaining the R-K difference even when the source task is difficult once the subjects learn how to make R-K judgment according to source memory component as shown in the results when fitting Exp 3-col R-K judgements with components learned from Exp 1.

Although false alarmed lure items and missed old items are not investigated in the current study, we hope this model could be further validated in different studies and on different datasets.

7.5 Chapter Acknowledgements

Chapter 7, in part, is the reorganized version of the material prepared for submission for publication. Kueida Liao; Matthew V. Mollison; Tim Curran; Virginia R. de Sa. "Using single-trial EEG classifiers to decompose remember-know difference", *in preparation*. The dissertation author is the primary author of this manuscript.

Chapter 8

Summary

In Chapter 3, we showed that subjects' trial-by-trial memory behaviors could be predicted during attempted memory retrieval using leave-one-trial-out (LOTO) pattern classifiers trained on their own data. In addition, the multivariate classification also showed different ERP components that were important to location source information retrieval and to color source information retrieval.

Based on the progress in Chapter 3, in Chapter 4, we further performed memory behavior prediction on memory EEG during memory retrieval using leave-one-subject-out (LOSO) classifiers trained on EEG data from different subjects. The success of such a subject-independent LOSO classifier showed that ERP features for memory behaviors during memory retrieval were consistent across subjects. Moreover, behaviors that originally were not with enough trials for an individual subject for training a classifier were possibly to be investigated using LOSO classifier because training data could come from multiple subjects and would no longer be insufficient.

In Chapter 5, we tried using LOSO pattern classifier to understand what were the ERP components associated with the familiarity responses in the modified R-K judgments in EEG data from recognition memory experiments . We found directly training a LOSO classifier to distinguish familiar responses from correct rejection responses using EEG data would have resulted in a classifier trained on both familiarity strength and confidence instead of pure familiarity strength. In order to correct the confidence bias in the training data, we developed a

confidence classifier designed to classify decision confidence on new trials and showed that the confidence classifier was not trained on memory strength. We then used this classifier to correct the confidence bias in the familiarity classifier and showed that remember and know responses were not different in terms of familiarity strength.

Based on the findings in Chapter 5, in Chapter 6, we defined single-trial LOSO EEG classifiers to represent three components (confidence, source memory, and item memory) that could be used to explain the difference between remember responses and familiar responses in R-K paradigm. Moreover, we showed that these three classifiers (components) could differentiate remember responses from familiarity responses. And then, we showed that the difference between remember responses and know responses could be interpreted as a combination of source memory, item memory, and confidence using LOSO linear regression model.

From the perspective of contributions to memory EEG analysis, we firstly showed that memory behavior could be successfully predicted using single-trial EEG pattern classifiers trained on ERP components consistent across subjects in Chapter 3 and 4. Later, in Chapter 5, we showed that the we could train a classifier more independently of the bias in the training data with bias control using the projections from the classifier associated with the bias. Finally, in Chapter 6, we demonstrated that the difference between complex behavior-pair could be decomposed into simple underlying components using single-trial EEG classifiers and a linear regression model.

In terms of the contributions to memory study, in Chapter 5 and 6, we showed that the confidence classifier was not trained on memory strength, which also indicated that confidence itself could be an underlying component during memory retrieval that was independent of memory strength. Last but not the least, we proposed a new three-component model in Chapter 7 based on our findings in both Chapter 5 and 6.

Bibliography

Abdulkader, S. N., Atia, A., & Mostafa, M.-S. M. (2015). Brain computer interfacing: Applications and challenges. *Egyptian Informatics Journal*, *16*(2), 213–230.

Addante, R. J., Ranganath, C., & Yonelinas, A. P. (2012). Examining ERP correlates of recognition memory: Evidence of accurate source recognition without recollection. *NeuroImage*, *62*(1), 439–450.

Agresti, A., & Caffo, B. (2000). Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures. *The American Statistician*, *54*(4), 280–288.

Anderson, J. R., Zhang, Q., Borst, J. P., & Walsh, M. M. (2016). The discovery of processing stages: Extension of sternberg's method. *Psychological Review*, *123*(5), 481.

Blankertz, B., Lemm, S., Treder, M., Haufe, S., & Müller, K.-R. (2011). Single-trial analysis and classification of ERP components—a tutorial. *NeuroImage*, *56*(2), 814–825.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.

Busey, T. A., Tunnicliff, J., Loftus, G. R., & Loftus, E. F. (2000). Accounts of the confidence-accuracy relation in recognition memory. *Psychonomic bulletin & review*, *7*(1), 26–48.

Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory & Cognition*, *28*(6), 923–938.

Curran, T. (2004). Effects of attention and confidence on the hypothesized ERP correlates of recollection and familiarity. *Neuropsychologia*, *42*(8), 1088–1106.

Curran, T., & Cleary, A. M. (2003). Using ERPs to dissociate recollection from familiarity in picture recognition. *Cognitive Brain Research*, *15*(2), 191–205.

Curran, T., & Hancock, J. (2007). The fn400 indexes familiarity-based recognition of faces. *NeuroImage*, *36*(2), 464–471.

Dien, J., Beal, D. J., & Berg, P. (2005). Optimizing principal components analysis of event-related potentials: matrix type, factor loading weighting, extraction, and rotations. *Clinical Neurophysiology*, *116*(8), 1808–1825.

Donaldson, D. I., & Rugg, M. D. (1998). Recognition memory for new associations: Electrophysiological evidence for the role of recollection. *Neuropsychologia*, *36*(5), 377–395.

Donaldson, W. (1996). The role of decision processes in remembering and knowing. *Memory & Cognition*, *24*(4), 523–533.

Dunn, J. C. (2004). Remember-know: a matter of confidence. *Psychological Review*, *111*(2), 524.

Fazli, S., Grozea, C., Danóczy, M., Blankertz, B., Popescu, F., & Müller, K.-R. (2009). Subject independent EEG-based BCI decoding. *Advances in Neural Information Processing Systems*, *22*, 513–521.

Finnigan, S., Humphreys, M. S., Dennis, S., & Geffen, G. (2002). ERP 'old/new'effects: memory strength and decisional factor (s). *Neuropsychologia*, *40*(13), 2288–2304.

Friedman, D., Cycowicz, Y. M., & Bersick, M. (2005). The late negative episodic memory effect: the effect of recapitulating study details at test. *Cognitive Brain Research*, *23*(2-3), 185–198.

Gardiner, J. M. (1988). Functional aspects of recollective experience. *Memory & Cognition*, *16*(4), 309–313.

Hammon, P. S., & de Sa, V. R. (2007). Preprocessing and meta-classification for brain-computer interfaces. *IEEE Transactions on Biomedical Engineering*, *54*(3), 518–525.

Hammon, P. S., Makeig, S., Poizner, H., Todorov, E., & De Sa, V. R. (2007). Predicting reaching targets from human EEG. *IEEE Signal Processing Magazine*, *25*(1), 69–77.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J. D., Blankertz, B., & Bießmann, F. (2014, feb). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110.

Herron, J. E. (2007). Decomposition of the ERP late posterior negativity: Effects of retrieval and response fluency. *Psychophysiology*, *44*(2), 233–244.

Hirshman, E., & Henzler, A. (1998). The role of decision processes in conscious recollection. *Psychological Science*, *9*(1), 61–65.

Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence.

Jafarpour, A., Fuentemilla, L., Horner, A. J., Penny, W., & Duzel, E. (2014). Replay of very early encoding representations during recollection. *Journal of Neuroscience*, *34*(1), 242–248.

Jafarpour, A., Horner, A., Fuentemilla, L., Penny, W., & Duzel, E. (2013). Decoding oscillatory representations and mechanisms in memory. *Neuropsychologia*, *51*(4), 772–780.

Johansson, M., & Mecklinger, A. (2003). The late posterior negativity in ERP studies of episodic memory: action monitoring and retrieval of attribute conjunctions. *Biological psychology*, *64*(1-2), 91–117.

Johnson, J. D., Price, M. H., & Leiker, E. K. (2015). Episodic retrieval involves early and sustained effects of reactivating information from encoding. *NeuroImage*, *106*, 300–310.

Kerrén, C., Linde-Domingo, J., & Hanslmayr, S. (2018). An optimal oscillatory phase for pattern reactivation during memory retrieval. *Current Biology*, *28*(21), 3383–3392.

Kim, D., Jeong, W., Kim, J. S., & Chung, C. K. (2020). Single-trial EEG connectivity of default mode network before and during encoding predicts subsequent memory outcome. *Frontiers in Systems Neuroscience*, *14*.

Klimesch, W. (1997). EEG-alpha rhythms and memory processes. *International Journal of psychophysiology*, *26*(1-3), 319–340.

Krauledat, M., Schröder, M., Blankertz, B., & Müller, K.-R. (2007). Reducing calibration time for brain-computer interfaces: A clustering approach. In *Advances in neural information processing systems* (pp. 753–760).

Ledoit, O., & Wolf, M. (2004). Honey, i shrunk the sample covariance matrix. *The Journal of Portfolio Management*, *30*(4), 110–119.

Leynes, P. A., & Phillips, M. C. (2008). Event-related potential (ERP) evidence for varied recollection during source monitoring. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(4), 741.

Liao, K., Mollison, M. V., Curran, T., & de Sa, V. R. (2018). Single-trial EEG predicts memory retrieval using leave-one-subject-out classification. In *2018 ieee international conference on bioinformatics and biomedicine (bibm)* (pp. 2613–2620).

Liao, K., Mollison, M. V., Curran, T., & de Sa, V. R. (2021). EEG reveals familiarity by controlling confidence in memory retrieval. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., & Arnaldi, B. (2007). A review of classification algorithms for EEG-based brain-computer interfaces. *Journal of neural engineering*, *4*(2), R1.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157.

Makeig, S., Onton, J., et al. (2011). ERP features and EEG dynamics: an ICA perspective. *Oxford Handbook of Event-Related Potential Components*.

Maris, E., & Oostenveld, R. (2007, aug). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190.

Mecklinger, A. (2006). Electrophysiological measures of familiarity memory. *Clinical EEG and Neuroscience*, *37*(4), 292–299.

Mollison, M. V., & Curran, T. (2012, sep). Familiarity in source memory. *Neuropsychologia*, *50*(11), 2546–2565.

Mousavi, M., Koerner, A. S., Zhang, Q., Noh, E., & de Sa, V. R. (2017). Improving motor imagery BCI with user response to feedback. *Brain-Computer Interfaces*, *4*(1-2), 74–86.

Müller-Putz, G., Scherer, R., Brunner, C., Leeb, R., & Pfurtscheller, G. (2008). Better than random: a closer look on BCI results. *International Journal of Bioelectromagnetism*, *10*, 52–55.

Nessler, D., Mecklinger, A., & Penney, T. B. (2001). Event related brain potentials and illusory memories: the effects of differential encoding. *Cognitive Brain Research*, *10*(3), 283–301.

Noh, E., & de Sa, V. (2014). Discriminative dimensionality reduction for analyzing EEG data. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 36).

Noh, E., Herzmann, G., Curran, T., & de Sa, V. R. (2014). Using single-trial EEG to predict and analyze subsequent memory. *NeuroImage*, *84*, 712–723.

Noh, E., Liao, K., Mollison, M. V., Curran, T., & Sa, V. R. d. (2018). Single-trial EEG analysis predicts memory retrieval and reveals source-dependent differences. *Frontiers in Human Neuroscience*, *12*, 258.

O'Sullivan, J. A., Power, A. J., Mesgarani, N., Rajaram, S., Foxe, J. J., Shinn-Cunningham, B. G., ... Lalor, E. C. (2015). Attentional selection in a cocktail party environment can be decoded from single-trial EEG. *Cerebral Cortex*, *25*(7), 1697–1706.

Parra, L. C., Spence, C. D., Gerson, A. D., & Sajda, P. (2005). Recipes for the linear analysis of EEG. *NeuroImage*, *28*(2), 326–341.

Ratcliff, R., Sederberg, P. B., Smith, T. A., & Childers, R. (2016). A single trial analysis of EEG in recognition memory: Tracking the neural correlates of memory strength. *Neuropsychologia*, *93*, 128–141.

Ray, A. M., Sitaram, R., Rana, M., Pasqualotto, E., Buyukturkoglu, K., Guan, C., ... Ruiz, S. (2015). A subject-independent pattern-based brain-computer interface. *Frontiers in Behavioral Neuroscience*, *9*, 269.

Rosburg, T., Mecklinger, A., & Frings, C. (2011). When the brain decides: a familiarity-based approach to the recognition heuristic as evidenced by event-related brain potentials. *Psychological Science*, *22*(12), 1527–1534.

Rubin, S. R., Petten, C. V., Glisky, E. L., & Newberg, W. M. (1999). Memory conjunction errors in younger and older adults: Event-related potential and neuropsychological data. *Cognitive Neuropsychology*, *16*(3-5), 459–488.

Rugg, M. D., & Curran, T. (2007). Event-related potentials and recognition memory. *Trends in Cognitive Sciences*, *11*(6), 251–257.

Rugg, M. D., Mark, R. E., Walla, P., Schloerscheidt, A. M., Birch, C. S., & Allan, K. (1998). Dissociation of the neural correlates of implicit and explicit memory. *Nature*, *392*(6676), 595–598.

Rutishauser, U., Aflalo, T., Rosario, E. R., Pouratian, N., & Andersen, R. A. (2018). Single-neuron representation of memory strength and recognition confidence in left human posterior parietal cortex. *Neuron*, *97*(1), 209–220.

Sarah, S. Y., & Rugg, M. D. (2010). Dissociation of the electrophysiological correlates of familiarity strength and item repetition. *Brain Research*, *1320*, 74–84.

Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, *4*(1).

Schwarze, U., Bingel, U., Badre, D., & Sommer, T. (2013). Ventral striatal activity correlates with memory confidence for old-and new-responses in a difficult recognition test. *PLOS One*, *8*(3), e54324.

Sun, X., Qian, C., Chen, Z., Wu, Z., Luo, B., & Pan, G. (2016). Remembered or forgotten?—an EEG-based computational prediction approach. *PLOS One*, *11*(12), e0167497.

Tsivilis, D., Allan, K., Roberts, J., Williams, N., Downes, J. J., & El-Deredy, W. (2015). Old-new ERP effects and remote memories: the late parietal effect is absent as recollection fails whereas the early mid-frontal effect persists as familiarity is retained. *Frontiers in Human Neuroscience*, *9*, 532.

Tulving, E. (1985). Memory and consciousness. *Canadian Psychology/Psychologie Canadienne*, *26*(1), 1.

Velu, P., & de Sa, V. R. (2013). Single-trial classification of gait and point movement preparation from human EEG. *Frontiers in Neuroscience*, *7*, 84.

Vilberg, K. L., Moosavi, R. F., & Rugg, M. D. (2006). The relationship between electrophysiological correlates of recollection and amount of information retrieved. *Brain Research*, *1122*(1), 161–170.

Voss, J. L., & Paller, K. A. (2008). Brain substrates of implicit and explicit memory: The importance of concurrently acquired neural signals of both memory types. *Neuropsychologia*, *46*(13), 3021–3029.

Wilcoxon, F. (1945). Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, *1*(6), 80–83.

Wilding, E. L. (2000). In what way does the parietal ERP old/new effect index recollection? *International Journal of Psychophysiology*, *35*(1), 81–87.

Wilding, E. L., Doyle, M. C., & Rugg, M. D. (1995). Recognition memory with and without retrieval of context: An event-related potential study. *Neuropsychologia*, *33*(6), 743–767.

Wilding, E. L., & Rugg, M. D. (1996). An event-related potential study of recognition memory with and without retrieval of source. *Brain*, *119*(3), 889–905.

Wilding, E. L., & Rugg, M. D. (1997). Event-related potentials and the recognition memory

exclusion task. *Neuropsychologia*, *35*(2), 119–128.

Wixted, J. T. (2007). Dual-process theory and signal-detection theory of recognition memory. *Psychological Review*, *114*(1), 152.

Wixted, J. T., & Mickes, L. (2010). A continuous dual-process model of remember/know judgments. *Psychological Review*, *117*(4), 1025.

Wixted, J. T., & Stretch, V. (2004). In defense of the signal detection interpretation of remember/know judgments. *Psychonomic Bulletin & Review*, *11*(4), 616–641.

Wolk, D. A., Schacter, D. L., Lygizos, M., Sen, N. M., Holcomb, P. J., Daffner, K. R., & Budson, A. E. (2006). ERP correlates of recognition memory: Effects of retention interval and false alarms. *Brain Research*, *1096*(1), 148–162.

Woodruff, C. C., Hayama, H. R., & Rugg, M. D. (2006). Electrophysiological dissociation of the neural correlates of recollection and familiarity. *Brain Research*, *1100*(1), 125–135.

Woroch, B., & Gonsalves, B. D. (2010). Event-related potential correlates of item and source memory strength. *Brain Research*, *1317*, 180–191.

Wynn, S. C., Daselaar, S. M., Kessels, R. P., & Schutter, D. J. (2019). The electrophysiology of subjectively perceived memory confidence in relation to recollection and familiarity. *Brain and Cognition*, *130*, 20–27.

Wynn, S. C., Kessels, R. P., & Schutter, D. J. (2020). Effects of parietal exogenous oscillatory field potentials on subjectively perceived memory confidence. *Neurobiology of Learning and Memory*, *168*, 107140.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(6), 1341.

Yonelinas, A. P. (1997). Recognition memory rocs for item and associative information: The contribution of recollection and familiarity. *Memory & Cognition*, *25*(6), 747–763.

Yonelinas, A. P., Kroll, N. E., Quamme, J. R., Lazzara, M. M., Sauvé, M.-J., Widaman, K. F., & Knight, R. T. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection and familiarity. *Nature Neuroscience*, *5*(11), 1236–1241.