The Time and Location of Natural Reading Processes in the Brain

Leila Wehbe

August 2015
CMU-ML-15-104

Machine Learning Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA

Thesis Committee:
Tom Mitchell, Chair
Eduard Hovy
Cosma Shalizi
Jack Gallant (University of California, Berkeley)
Brian Murphy (Queen's University Belfast)

Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Copyright © 2015 Leila Wehbe

This research was sponsored by: National Institutes of Health grant numbers R90DA023420 and R01HD075328; National Science Foundation grant number IIS0835797; Air Force Research Laboratory contract number FA865013C7360; and W.M. Keck Foundation grant number DT123107.

Keywords: Functional NeuroImaging, Predicting Brain Activity, Language Processing, FMRI, MEG, Naturalistic Experiments, Story Reading, Hypothesis Testing, Conditional Independence Testing

Abstract

How is information organized in the brain during natural reading? Where and when do the required processes occur, such as the perception of individual words and the construction of sentence meanings. How are semantics, syntax and higher-level narrative structure represented? Answering these questions is core to understanding how the brain processes language and organizes complex information. However, due to the complexity of language processing, most brain imaging studies focus only on one of these questions using highly controlled stimuli which may not generalize beyond the experimental setting.

This thesis proposes an alternative framework to study language processing. We acquire data using a naturalistic reading paradigm, annotate the presented text using natural language processing tools and predict brain activity with machine learning techniques. Finally, statistical testing is used to form rigorous conclusions. We also suggest the use of direct non-parametric hypothesis tests that do not rely on any model assumptions, and therefore do not suffer from model misspecification.

Using our framework, we construct a brain reading map from functional magnetic resonance imaging data of subjects reading a chapter of a popular book. This map represents regions that our model reveals to be representing syntactic, semantic, visual and narrative information. Using this single experiment, our approach replicates many results from a wide range of classical studies that each focus on one aspect of language processing.

We extend our brain reading map to include temporal dynamics as well as spatial information by using magnetoencephalography. We obtain a spatio-temporal picture of how successive words are processed by the brain. We show the progressive perception of each word in a posterior to anterior fashion. For each region along this pathway we show a differentiation of the word properties that best explain its activity.

Acknowledgments

This thesis and the ideas it is based on could not have been possible without Tom Mitchell. A brainstorming session with Tom is a unique experience, during which many new perspectives are born. I am very thankful for the freedom that Tom gave me to explore the directions I was interested in and for his support and belief in them.

This work also relies on the help of many people in our lab. Gustavo Sudre was very helpful when I first started. I have enjoyed going through the long PhD program in parallel with Alona Fyshe, taking classes, going through milestones and collaborating together. Nicole Rafidi was helpful in countless discussions, and her friendship has offered a lot of support in many other ways. I have learned a lot from Brian Murphy who gave me very good advice on many occasions. Partha Talukdar was helpful at multiple times in proposing new methods. I thank Erika Laing for collecting MEG data for many subjects, and Dan Howarth for preprocessing that data and answering my many questions. I also thank Kai-min Chang, Mark Palatucci, Dean Pomerleau, and more recent members such as Dan Schwartz and Mariya Toneva, as well as all the previously mentioned members of the lab, for all the mind opening lab meetings we have had in which I have learned much more than I can realize.

I would like to also thank many other fantastic people from outside the lab, including my collaborators: Ashish Vaswani and Kevin Knight for their enthusiasm to engage in novel directions, Rebecca Steorts for her great ideas and her dedication and Cosma Shalizi for his great insight and tremendous help. I have received invaluable advice from Marcel Just, Mike Tarr, Jack Gallant and Eduard Hovy. I am very grateful to Nancy Kanwisher for her support and guidance since I was an undergrad. I am also grateful to Evelina Fedorenko for great discussions. I would like to thank Scott Kurdilla and Deborah Viszlay for helping me acquire fMRI data, and finally John Pyles and Vladimir Cherkassky for helping me learn how to preprocess it.

The machine learning department in CMU is a very unique place, with very strong social cohesion. Everyone is a friend and you can always find support. I thank Diane Stidle for helping to make it so, and for her assistance on many occasions throughout the years. I thank all my friends in the department, including Willie Neiswanger for his energy and solidarity, Anthony Platanios for his positivity, Lucia Castellanos for the great life conversations and Kirstin Early for her thoughtfulness. Avinava Dubey has always been very helpful with his uplifting spirit. Will Bishop has been a great writing partner with his incredible generosity. I thank Aaditya Ramdas for his energy and resourcefulness, for being a very helpful friend as well as a very meticulous and engaged collaborator. I also thank my oldest grad school friends: Madalina Fiterau, Ankur Parikh and Min Xu, it has been great to go through the entire experience together and reflect upon it during our late-night thoughtful discussions. Finally, I thank my oldest friend, Jessica Chemali, for sharing my interest in research and in the brain ever since college, for all her help, and for still inspiring me with her enthusiastic, positive spirit that can withstand any hard times.

Lastly, I would like to thank my parents and my brother for their understanding and their support in this long endeavor, so far away from home.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Reading a story is a highly complex cognitive task that combines the low level perception of individual words, the representation of their meanings and parts of speech, the understanding of the grammar and meaning of entire sentences, and finally the tying of these individual sentences together into a coherent understanding of the story plot and the evolving beliefs, desires, emotions, and actions of story characters.

The vast majority of existing functional neuroimaging studies of reading and language processing are what we call *controlled experiments*. They are of a reductionist nature: they are carefully designed to isolate a specific task while controlling for all other variables. They thus allow the experimenter to make precise conclusions about the brain representation of that task. For instance, an experimenter might look for the brain regions responsible for understanding the structure of a sentence by contrasting the brain activity when sentences are read with the activity related to reading word-lists.

While controlled experiments are required in order to make specific, testable conclusions about brain function, they are not sufficient for understanding how complex processes are performed [51]. For example, it has been difficult for the field of language processing to converge on a single model of how the brain perceives and understands language [28, 47, 55]. This is due at least in part to the difficulty of understanding how such a complex system works by isolating one of its subprocesses at a time. This concept was mentioned in Allen Newell's 1973 essay *You can't play 20 questions with nature and win* [89], in which he summarizes the difficulty of combining the results of a large set of cognitive science experiments, and suggest a better alternative is to choose "a single complex task and do all of it". This concept has also been made atemporal by the fable of the *blind men and an elephant*, in which blind men fail to converge on a perception of an elephant after each of them perceives only one body part of the elephant.

In this thesis, we instead advocate an increased use of *naturalistic experiments* as an invaluable tool in the arsenal of the cognitive neuroscientist. In these experiments, the brain is imaged while it performs a natural task that mimics its real life behavior, and the complex interaction between the task's subprocesses can be studied. Some researchers already use naturalistic stimuli, such as natural videos [90], math problems [2] or natural stories [114], but this is still rather uncommon. We do not think that naturalistic experiments are a replacement for controlled experiments, but are rather complementary to them; they can study high-level processes and generate much needed hypotheses which could later be tested with targeted controlled experiments.

This thesis presents an integrated approach to study reading and language processing in the brain using naturalistic experiments, and includes first results on the timing and the location of different language subprocesses during natural reading. This thesis is therefore a methodological effort that is accompanied by novel results, and spans many disciplines:

1. **Cognitive neuroscience of language:** We present an approach to study the brain in its natural behavior, using a real text that we present without repetitions to the subject. Using functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG), we present interesting results about the rich information processing in the brain during language processing.
2. **Statistical Machine Learning:** To study natural reading, this thesis relies on finding relationships between the brain activity at different locations and timings and different properties of the rich stimulus text being read. We present a complex investigation pipeline that we use to provide unprecedentedly rich spatiotemporal brain maps indicating how language sub-processes are represented. We continuously and closely inspect and update our methods to increase the robustness of our results.
3. **Natural Language Processing (NLP):** While it does not present new work in NLP, this thesis uses many NLP tools in order to represent the content of the stimulus text in an intermediate space that describes its different properties. While we use specific models, which we think are reasonable to represent this content, our approach is agnostic to the model that is chosen. The interested researcher can consider other models they find more appropriate to their hypotheses or aims, and compare their ability to predict brain activity (while being cautious to follow proper statistical methodology).

The study of language processing

Story understanding and language processing have long been central topics of study across diverse fields including linguistics, computer science [72], cognitive science [74], literature and philosophy [17]. Today, a network of multiple brain regions are considered to be implicated in language [16, 21]. While the field started with a simplistic dissociation between the roles of Broca's area and Wernicke's area, current theories about language comprehension are more complex and most of them involve different streams of information that involve multiple regions (including Broca's and Wernicke's).

One of the main questions that still occupies the field is to understand the role of these multiple regions in response to language processing. Most experimental brain imaging studies of language processing have tackled the inherent complexity of the phenomenon by focusing on just one aspect of language at a time, via carefully controlled experiments. These experiments usually aim to find where in the brain this language process is located, or when after stimulus onset it occurs. For example, researchers have searched for brain regions where neural activity increases or decreases when the input stimulus is a word, in contrast to a non-word letter string [108]; or for a sentence with simple versus complex syntax [16]; or for a sentence with expected versus unexpected meaning [67]. These experiments require carefully controlled, hand-tailored textual stimuli that vary solely along one dimension of interest, raising the question of how much these findings reflect language processing in complex every-day use. For instance, the experimental

task can be too simplified or contrived that it will not require genuine linguistic processing, and could be solved using other high level cognitive abilities such as strategizing. Moreover, the stimuli for such experiments are often of a small number of types, each of which repeated again and again, effectively sampling a small linguistic space and reducing the generalization power of what is learned about the brain.

As a result of the different experimental setups that are used and the different focused hypotheses that are tested, different theoretical models have emerged. Accordingly, this has led to little agreement in the field, including on fundamental questions such as: Are language regions *specific* to language, or are they used for other functions? [21]. The disagreement concerns other questions as well, such as the role of the different language regions and the differentiation between regions processing syntax and regions processing semantics. Different models of meaning integration have been also proposed that disagree on the order in which semantic and syntactic information is accessed as a word is encountered, as well as on the order of integration of this information [26, 46].

The discrepancies in the field might be due to more than just methodological differences. They might be due to the difficulty of constructing a picture of how a complex system like language processing works by isolating the parts. In order to make trustworthy causal inferences, controlled experiments are a must, as they allow the experimenter to filter out the effect of confounding effects and variables. However, it seems very difficult to arrive at an understanding of how the processes of a complex system work together by isolating one process at a time, keeping everything else constant and varying only this process- a solution could be to instead focus on a complex task and study all the processes together [89].

This thesis attempts to support future consensus in the field, by proposing a rich model of language processing that explains the activity when language is processed in natural settings. This model would explicitly take into account the content of the stimulus that varies along many dimensions, and understand how different regions of the brain are related to these different dimensions. We present here our attempt to construct such a model, as well as encouraging results. These results are to be taken as a first step of a long line of investigation. After eventually establishing a higher level picture of language processing that one can trust, this empirical, data-driven model can later be tested with targeted controlled experiments.

Naturalistic imaging methodology

Our approach relies on using naturalistic stimuli: real texts that vary along many different dimensions simultaneously. Our approach involves building a rich model of language processing that explicitly represents the content of the language stimulus and the processing needs it requires in an intermediate feature space. This feature space can contain very diverse information about the semantic properties of the text, the syntactic structure, the narrative structure etc. We examine, using multiple methods, which brain areas have activity that is modulated by the different types of information, leading us to distinguish between brain areas on the basis of which type of information they represent, and the latency within which they process that information.

The methods presented in this thesis can be used as a unified framework to test and contrast competing theories of reading and story understanding. As long as different theories can be characterized in terms of different time series of annotated story features, our approach can compare

them by training on these alternative feature sets, then testing experimentally which theory offers a better prediction of brain data beyond the training set. Our approach differs in multiple key respects from typical language studies. First, the subjects in our study read an authentic book chapter, exposing them to the rich lexical and syntactic variety of a non-constructed text that evokes a natural distribution of the many neural processes involved in diverse, real-world language processing. Second, our analysis method differs significantly from studies that search for brain regions where the magnitude of neural activity increases along one stimulus dimension. Instead, we try to find what is being encoded in brain activity.

Analysis methodology

The following are more details about the approach. We train a comprehensive generative model that incorporates the effects of many different aspects of language processing. Given a text passage as input, this trained computational model outputs a time series of fMRI or MEG activity that it predicts will be observed when the subject reads that passage. The text passage input to the model is annotated with a set of detailed features for each word, representing a wide range of language features: from the number of letters in the individual word, to its part of speech, to its role in the parse of its sentence, to a summary of the emotions and events involving different story characters. The model makes predictions of the brain activation for the text passage by capturing how this diverse set of information contributes to the brain activity. The model can therefore make testable predictions of the brain activity associated with novel text passages, which may vary arbitrary in their content. Our model accounts for the different levels of processing involved in story comprehension; however, it doesn't stop at modeling the presence/absence of a story process, such as the presence of story characters, or the presence of emotional content. It can go even further by explicitly searching for the brain activity represenations for individual stimuli such as the mention of a specific story character, the presence of a specific emotion and the use of a specific syntactic part-of-speech or the occurrence of a given semantic feature.

Outline

In chapter 2, we present the relevant literature and survey language models and relevant computational methods that are related to our approach.

In chapter 3 we present our own approach in detail[1], and we present different learning models that we have used to try and improve it[2].

In chapter 4 we describe an fMRI natural reading experiment and the resulting brain representation maps we were able to obtain from it[3]. These maps indicate where different reading processes (semantic, syntactic, visual and narrative properties) are processed in the brain as suggested by our model. Our model therefore not only recovers areas implicated in language processing but also differentiates between them on the basis of the language properties they appear to be representing. Using one reading experiment, this model is able to replicate results from a

[1] Joint work with Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas and Tom Mitchell, [127].
[2] Joint work with Aaditya Ramdas, Rebecca Steorts and Cosma Shalizi [126].
[3] Joint work with Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas and Tom Mitchell [127].

variety of experiments that are targeted towards one aspect of language, as well as suggest new hypothesis about the brain representation of language.

In chapter 5, we apply our investigation methods to an MEG experiment[4]. We use neural network models of language in order to obtain a numerical assessment of the context in which each word occurs in the story. We suggest a classification test that aims at testing for conditional independence. Conditional independence is needed to assess if a text feature is being represented in an area, or if it appears to be because it is correlated with another feature that is itself represented in this area. We use this test to obtain a spatio-temporal picture of how consecutive words are perceived by the brain, which includes results about how different word features are perceived at different times and locations.

In chapter 6 we present insight about how many brain image analysis methods, including ours, constitute indirect hypothesis tests, and suggest the use of direct hypothesis tests[5]. We illustrate these methods with examples that suggest that they might give more interpretable results.

We end the thesis with a summary of our contributions; they provide a proof-of-concept for our approach and are also a first step in uncovering how the brain represents the meaning of natural text.

[4]Joint work with Ashish Vaswani, Kevin Knight and Tom Mitchell [128], as well as Dan Howarth and Erika Laing.

[5]Joint work with Aaditya Ramdas and Tom Mitchell.

Chapter 2

Related work

This thesis builds upon advances in three different lines of investigation: (1) cognitive neuroscience of language processing, (2) computational approaches for functional neuroimaging and (3) the more recent introduction of naturalistic stimuli to the study of brain functions. We survey in this chapter some of the major advances and some of the most recent contributions.

2.1 Neurobiology of language

The field of neurobiology of language started with lesion studies. Because patients with lesions to the area around the anterior superior temporal cortex had trouble understanding language, and patients with legions to Brodmann Areas (BA) 44-45 presented with troubles in speech production, the classical model of the neurobiology of language considers these two areas as the main language territory in the brain. The first area is know as Wernicke's area and is considered responsible for language understanding, while the second is known as Broca's area and is considered to be responsible for speech production under the classical model. Fig. 2.1 shows this classical model.

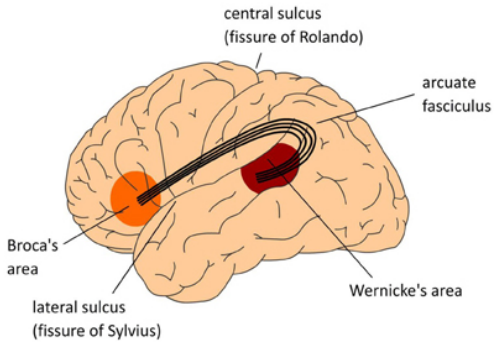


Figure 2.1: The classical Wernicke-Lichtheim-Geschwind model of the neurobiology of language. In this model Broca's area is crucial for language production, Wernicke's area subserves language comprehension, and the necessary information exchange between these areas (such as in reading aloud) is done via the arcuate fasciculus, a major fiber bundle connecting the language areas in temporal cortex (Wernicke's area) and frontal cortex (Broca's area). The language areas are bordering one of the major fissures in the brain, the so-called Sylvian fissure. Collectively, this part of the brain is often referred to as perisylvian cortex. *This figure was taken from [47].*

As non-invasive brain imaging methods were developed they allowed much more extensive research on the brain basis of language. Multiple brain imaging methods have been used to

study language, such as Electroencephalography (EEG), Functional Magnetic Resonance Imaging (fMRI) and Magnetoencephalography (MEG). Studies using EEG have been very common to study different evoked responses due to various types of language processes. EEG measures the change in the voltage on the scalp due to neural activity. The signal is therefore directly related to neural activity and has no latency, but it has very bad spatial resolution because of the distortion caused by the scalp. MEG records the change in the magnetic field on the surface of the head that is caused by neural activity. MEG recordings are also directly related to neural activity and have no latency, and the signal is less distorted than EEG and has better spatial resolution. Finally, fMRI is a also very commonly used tool. FMRI records the change in the oxygen level of the blood due to neural activity. This change is slow (it takes around 6 seconds after onset for it to peak), and furthermore, it takes 1-2s to obtain one brain image. While it suffers from low temporal resolution, fMRI has however a great spatial resolution on the order of millimeters, and has therefore been very popular as a tool for looking at the spatial localization of language processes.

2.1.1 Single word processing

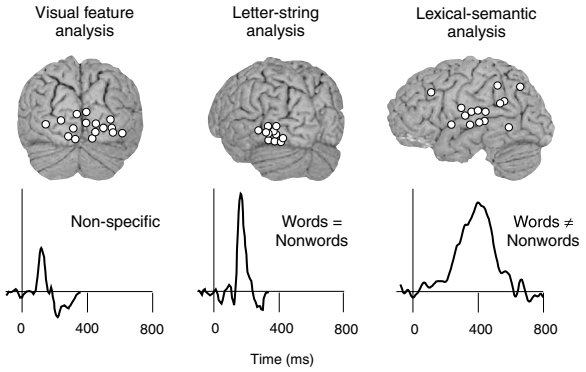


Figure 2.2: Cortical dynamics of silent reading. Dots represent projected sources of activity in the visual cortex (left brain sketch) and the temporal cortex (right brain sketch). The curves display the mean time course of activation in the depicted source areas for different conditions. The initial visual feature analysis in the visual cortex at ∼100 ms is non-specific to language. Comparing responses to letter strings and other visual stimuli reveals that letter string analysis occurs around 150 ms. Finally comparing the responses to words and non-words (made-up words) reveals lexical-semantic analysis in the temporal cortex at ∼200-500ms. *This figure was adapted from [108].*

Humans read with an average speed of 3 words per second. Reading requires us to perceive incoming words and gradually integrate them into a representation of the meaning. As words are read, it takes 100ms for the visual input to reach the visual cortex. 50ms later, the visual input is processed as letter strings in a specialized region of the left visual cortex [108]. Between 200-600ms, the word's semantic properties are processed (see Fig. 2.2). This is inferred because between 200-600ms, there is a sustained activity when reading words versus non-words [108].

The period 200-600ms is also known for a mismatch negativity response called the N400: the response is more negative for semantically incongruous words [56]. The N400 has been studied extensively, and, more recently, it has been shown to be a graded response that is a function of how surprising the word is, instead of an all-or-none event. In [25], the amount of surprisal of a word (given its context) is shown to predict the intensity of the N400 response (surprisal was computed using different kind of statistical language models).

The P600 is another highly studied mismatch in event related potential. Around 600ms, the activity is more positive when certain kinds of incongruities are present, of mainly a syntactic nature (e.g. a mistake in the conjugation of a verb) [67]. However, the reasons behind a P600 can be more complex. For example, a semantically congruous sentence which attributes to a non-animate object an animate action will produce a P600 and not a N400.

Word meaning

When the brain perceives a word, this activates areas associated with functions that are used in everyday life to interact with that word. For example, if the word is a verb denoting a physical action, the motor brain areas that are associated with that action have been shown to be activated [101]. Another example is that food related words can activate gustatory cortex [35]. This has created a large debate about whether the activity in these non-language regions are part of the meaning of the word (also known as the embodied meaning) or if they are mere brain connections brought about by many years of the language system and these different regions firing together when we think of the word while interacting with it [101].

[79] used the relationship between semantic properties and brain activity of different regions in order to decode the word a subject was seeing. The properties of that word were summarized by a vector indicating how often they are used with 25 hand picked verbs. This vector was therefore used as an approximation for each word's meaning. The brain activity of every voxel in the brain (volume pixel) was predicted as a function of these verb co-occurences (see Fig. 2.7). The authors were therefore able to model the relationship between the word's features and the brain activity in various parts of the brain. In [117], the authors follow a similar approach of representing the same nouns with a vector of hand labeled features, and find that these features are related to the signal in different regions at different points in time.

2.1.2 Sentence processing models

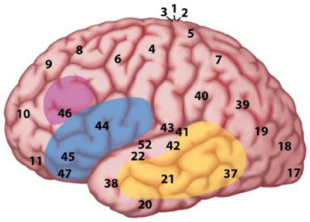


Figure 2.3: The MUC model of language. The figure displays a lateral view of the left hemisphere. The numbers indicate Brodmann areas. These are areas with differences in the cytoarchitectonics (i.e., composition of cell types). The memory areas are in the temporal cortex (in yellow) including the angular gyrus in parietal cortex. Unification requires the contribution of Broca's area (Brodmann areas 44 and 45) and adjacent cortex (Brodmann areas 47 and 6) in the frontal lobe. Control operations recruit another part of the frontal lobe (in pink), and the Anterior Cingulate Cortex (ACC; not shown in the figure), as well as areas involved in attention. *This figure was taken from [47]* .

Many models of sentence processing have been proposed in the literature, and they disagree significantly on the function of certain regions. Hagoort's Memory Unification Control (MUC) model of language processing designates the temporal cortex as the center for semantic memory, and Broca's area as a unification center that serves to combine different semantic concepts into novel ones [47]. Finally a control center in the frontal lobe and the anterior cingulate cortex is responsible for high level functions such as ones required for social interaction (see Fig. 2.3).

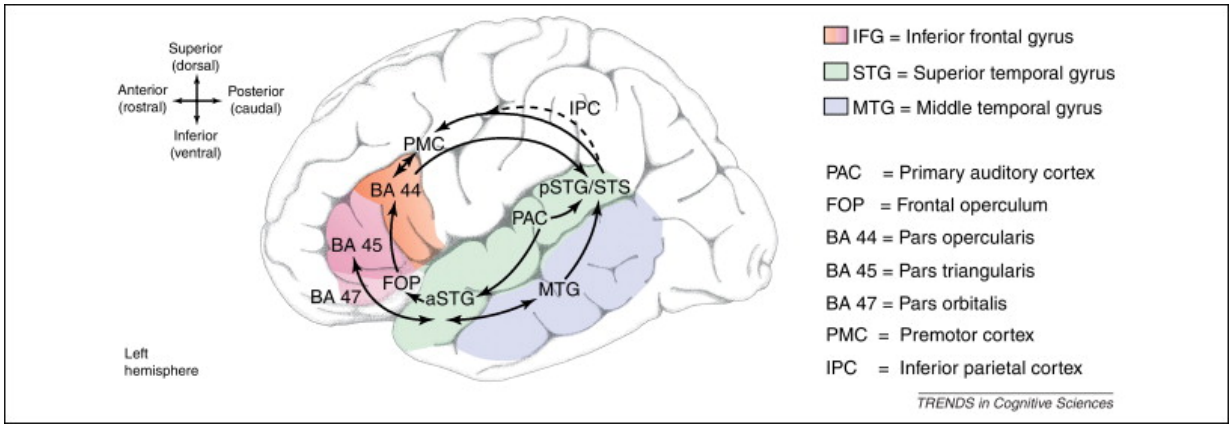


Figure 2.4: The cortical language circuit (schematic view of the left hemisphere). After the word is perceived in the primary auditory cortex, semantic information becomes available in MTG and syntactic information becomes available in the aSTG. Semantic information flows from the MTG to the anterior IFG (BA45, BA47) and syntactic information flows from the aSTG to the posterior IFG (BA44). Information then flows in a top-down fashion: the semantic information flows from the anterior IFG to the pSTG and syntactic information flows from the posterior IFG to the pSTG. *This figure was adapted from [28].*

In Friederici's cortical language circuit, after the word is perceived by the primary auditory cortex, syntactic and semantic information travel along distinct routes: syntactic information propagates through the anterior STG towards BA 44, while the semantic information travels from the MTG to BA 45 (see Fig. 2.4 for the diagram and the names of the regions). The next step is a top-down process that results in the semantic and syntactic information being sent to the posterior STG were they are combined.

In the Hickok and Poeppel dual-stream model [55], the lexical and combinatorial processes are constrained to the ventral stream (mainly located in the temporal cortex, see figure 2.5). The dorsal stream is responsible for the sensorimotor processes responsible for speech production.

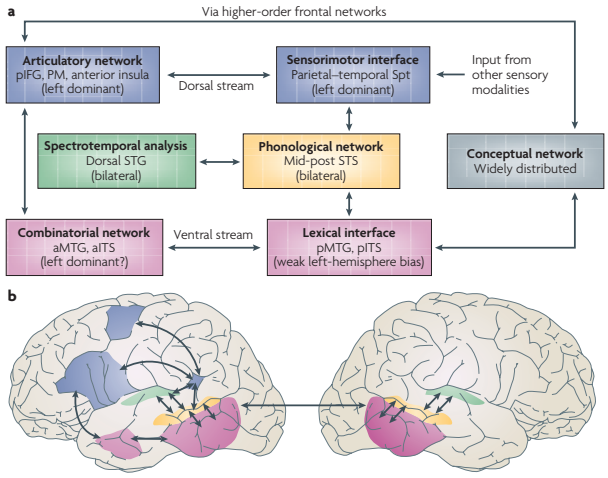


Figure 2.5: The dual-stream model of the functional anatomy of language. (a) Diagram of the dual-stream model. Speech processing starts early in the bilateral auditory cortices. The system diverges into a dorsal stream that maps phonological representations to articulatory motor representations and a ventral stream that maps phonological representations to lexical conceptual representations. (b) Approximate location of the brain region involved in different states of the model. *This figure was adapted from [55].*

We have seen in this section that the current theories about language comprehension in the brain are more complex than the originally proposed Wernicke-Lichtheim-Geschwind model of

the neurobiology of language. Most of the proposed models are comprised of different streams of information that involve multiple regions, including Broca's and Wenicke's areas. However, the models rarely agree. Indeed, the field has not come to a unified theory about such basic question as the differentiation between regions processing syntax and regions processing semantics. E.g. [22] has found no regions to be responsive exclusively to syntactic or semantic information, while [18] has found regions in the IFG that exclusively process syntax or semantics.

The disparity between these results might be due to the experimental setups that have been used, or the different analysis methods. It might also be due to the fact that language processing is an extremely complex task involving many subparts, and it is difficult to construct a picture of how the entire system is working when looking at one individual part of the system at a time. While designing a controlled experiment is the correct way to arrive to an inference one can trust, it is difficult to generalize how various parts of the system work together if we keep all of them fixed and vary one.

Therefore, we argue that a way to reduce the disagreement in the field could be a rich model of language processing that explicitly models the content of the language stimulus, and via this rich annotation of the text, this model would be able to find regions that are responsive to different properties of the text (such as semantics or syntax). Such a model might also be able to find the point in time when these regions are activated. This thesis is an attempt to produce such a model, and in chapters 4 and 5 we provide results that are a first step along this approach. In order to increase the confidence in the results of such an approach, they can be later supplemented by controlled experiments that test the various subparts.

Another highly debated question in the neurobiology of language is: Are language regions specific to language? In [24], a diagram of the language network is provided that accounts for the different definitions of that network (see Fig. 2.6). At the core of this network is a set of "high-level" language regions that is hypothesized to be functionally specialized for language, i.e. to be mainly responsible of language processing. The authors extend the traditional definition of a functionally specialized region to the notion of a functionally specialized network: the areas comprising that network might be implicated in other tasks, however the network itself is functionally specialized for one task.

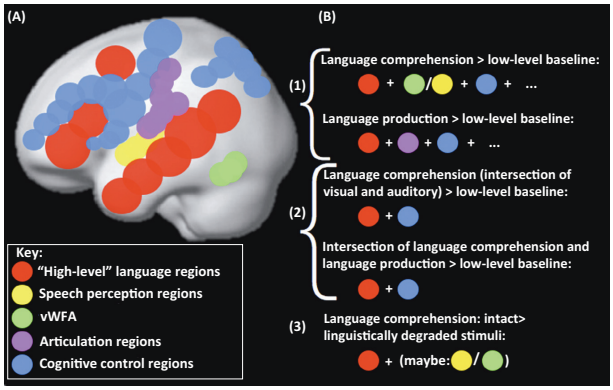


Figure 2.6: The language network under different definitions. (A) A schematic depiction of five sets of brain regions that are sometimes included in the language network: red, the classic high-level language-processing regions; yellow, speech perception regions; green, visual word-form area; purple, speech articulation regions; and blue, cognitive control regions. (B) A schematic illustration of possible definitions of the language network, ranging from very liberal (1) to more conservative (2 and 3). *This figure was taken from [24].*

2.2 Computational modeling of human brain activity

A very common method for analyzing fMRI data is a univariate analysis in which the activity of a voxel is modeled as a function of the few experimental conditions. Usually, a design matrix of the experiment is constructed that indicates when different stimuli occur[1]. Each voxel's activity is then regressed on this design matrix. By testing whether the regression parameters for each condition are significantly different, one can conclude if the voxel responds to a given condition significantly more than rest, or if it responds to one condition significantly more than another. Since this approach considers only one voxel at a time, it might not be sensitive enough to detect brain representations that manifest as a pattern of subtle changes in activity in a brain area. As a solution for this, many experimenters prefer to use brain decoding.

2.2.1 Brain decoders and generative models

Given the brain activity of a subject, a brain decoder is a classifier that predicts what stimulus the subject was processing [80]. After training on a subset of images, a brain decoder is faced with a new brain image, and it asked to guess whether it corresponds to brain state A or B (e.g. whether the subject is seeing a picture of a face or a house). Brain decoding is also known as multi-voxel pattern analysis (MVPA) [93], and it has increased in popularity in the past decade, and has been used to differentiate between a large variety of conditions and with classifiers with differing complexities. The classifier trains on a brain image (or a part of a brain image) and is able to learn multivariate patterns and thus can be more sensitive than classical univariate analyses.

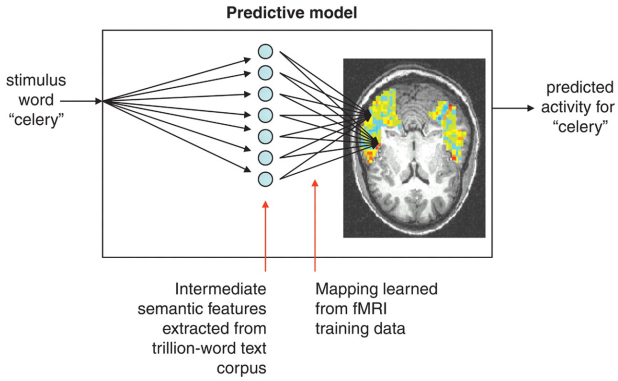


Figure 2.7: Form of the model for predicting fMRI activation for arbitrary noun stimuli. fMRI activation is predicted in a two-step process. The first step encodes the meaning of the input stimulus word in terms of intermediate semantic features whose values are extracted from a large corpus of text exhibiting typical word use. The second step predicts the fMRI image as a linear combination of the fMRI signatures associated with each of these intermediate semantic features. *This figure was taken from [79].*

There are many cases in which the experimenter is interested in complex stimulus that does not fit neatly into categories. For instance, one might be interested in studying the brain response to natural images or to the meanings of different words. In this case, a simple brain decoder that distinguishes between a small number of conditions might not be very useful. This is why some researchers use a generative model, or what is known as an encoding model [86]. Unlike decoders that express the stimulus as a function of the brain activity, encoding models express the brain activity as a function of the stimulus. When dealing with complex stimulus, one has to go through an intermediate step: building an Intermediate Feature Space (IFS) for the stimulus. An

[1]This matrix is convolved with a function designed to account for the delay in brain response. See [4].

IFS is an abstraction of the properties of the stimulus. For example, it can represent the semantic features of a word, or the visual features of an image, or the semantic features of an image. The brain activity is then predicted as a function of the IFS. For example, the authors of [79] express a voxel's activity as a linear combination of the semantic properties of the word the subject is reading (see Fig. 2.7), the parameters of which they learn from the brain data.

The IFS approach allows the experimenter to build a model that generalizes from a finite training set and predicts the activity for stimuli that were not seen in training. This is not possible for decoders that need samples from every condition to train on. Instead of learning the responses for individual stimuli like "house", "car", "table" etc., the IFS allows the experimenter to learn the brain responses for a set of concepts or features (e.g. "being manmade", "being a vehicle", "being made of wood"), and then predict the activity for new stimuli that have different configurations of these features. Because of this abstraction ability, IFS analysis is essential for the study of complex stimulus, and could help us figure out how the brain works by proposing models that explain its activity.

2.2.2 Corpus-based semantic models

In order to build appropriate IFS representations for the meanings of words, one can use vector space models of semantics. These models approximate the meaning of a word by assigning to it a vector of statistics computed from a corpus. In [79], the co-occurence of nouns with 25 verbs constitutes a vector space model that is used to predict brain activity. In [85], the authors compare different co-occurence statistics in terms of how well they perform on the brain prediction task. These models vary in the amount of syntactic structure they were using to compute co-occurence: the simplest model counts how many times words occur within 4 words of the target word. More complex models incorporate structure by including directionality[2], or even by including the entire parse tree of each sentence in the corpus[3]. The authors find that the most complex model (incorporating the parse tree) is the best at predicting the brain activity related to a word.

In order to be used to predict brain activity, the dimensionality of the corpus concurrence vector (on the order of the size of the dictionary, i.e. in the tens of thousands of entries) has to be reduced. This can be done using methods like principal component analysis, or other methods such as Non-Negative Sparse Embeddings [83] that enforced the learned representations to be sparse and non-negative, which leads to more interpretable dimensions. In [32], a vector space model is optimized by including the constraint of minimizing the error these vectors have when predicting brain activity. The resulting vectors perform better at predicting brain activity for new subjects and they also match a behavioral measure of semantics more closely than the vectors obtained without the brain data constraint. This experiment is a proof of concept that brain data can be utilized to improve statistical language models.

[2]i.e. how many times words occur on the left or right of the target word. Directionality is related to the grammatical relationship between words, e.g. objects are always on the right of their verbs.

[3]This model computes how many times words occur in specific relationships with the target word.

2.3 Naturalistic experiments

Naturalistic experiments that mimic real life human cognitive processing are an emerging trend that avoids the use of artificial stimuli, because they might not generalize well to real life conditions [51]. Recent experiments include subjects watching videos [90], solving math problems [2] and listening to stories [10].

2.3.1 Video Processing

In [54], subjects watch a movie in the scanner, and the brain activity in their visual cortex is transformed into a common space using a method the author call *hyperalignment*. The common space allows the authors to perform between-subject classification accurately, i.e. they are able to guess the part of the movie a subject is watching from a segment of their brain activity, given only the timeline of activity for the other subjects.

In [90], human subjects watching a series of short videos and an encoding model is built that predicts their brain activity as a function of the features of that video. This model is used in order to reconstruct what the subjects are seeing by predicting the brain activity for a large dataset of videos, picking the videos that are closest in prediction and averaging them in order to obtain a reconstruction of the original stimulus video. In [61], human subjects watch a series of short videos where a large set of objects were identified and annotated with semantic features. A mapping is constructed between the semantic features of the objects and the brain activity in different regions, and objects are revealed to be represented in a continuous semantic space that has smooth gradations over the visual and non-visual cortex.

2.3.2 Story Processing

Syntax

A few recent experiments have featured reading or listening to stories. In [5], subjects were scanned using fMRI while they listened to short stories that were written to specifically have various syntactic complexities with higher density than in usual text. Many syntactic complexity measures were computed, as well as other measures such as semantic surprisal and measures of Theory of Mind computations. Different brain regions, mainly in the temporal and inferior frontal cortices, are found to be correlated with these different measures.

In [10], subjects listened to a story in the fMRI scanner. The amount of syntactic structure analysis needed at each word was measured by building a parse tree of every sentence and computing the depth of each word, therefore assessing how complex the syntactic representation the subjects were processing at every word. This measure correlated with the activity in the anterior Temporal Lobe (aTL), indicating that the aTL is involved in computing syntactic structure under natural conditions.

Discourse processing

Other experiments have studied the processing of narrative structure while listening to stories. In [114], participants read short stories that were annotated by references to temporal information, changes in the causal relationships between narrated activities, points when the subject of the text changed, changes in characters, spatial locations, interactions with objects and and points when a character initiated a new goal. The authors found different regions in the brain that correlate in activity with the onset pattern of these events. For instance, the left and right superior temporal gyrii were correlated with changes of characters. In [74], a "protagonist's perspective interpreter network" is hypothesized to mainly be located in the right superior temporal cortex and the bilateral medial frontal cortex, based on a review of multiple studies. [74] also identifies a spatial imagery network in the bilateral intraparietal sulcus.

Chapter 3

Investigation Methods

This thesis is based on building a generative computational model for language processing that underlies natural story reading. We are interested in building a predictive model that expresses brain activity as a function of the content of what is being read. For that purpose, we construct different **intermediate feature spaces (IFS)**. These IFS express the properties of the text along a specific process. For example they can express the semantic or syntactic processes involved in understanding a text. They allow us to generalize our prediction to novel, unseen text. We use the model we learn as a investigation tool to reveal which regions of the brain are involved in a specific language process. Our working assumption is that if the model is able to correctly predict the brain activity in a given region using an IFS derived from a specific process, as measured by a classification task, then this suggests that the brain region might be involved in the given process. Figure 3.1 summarizes our approach.

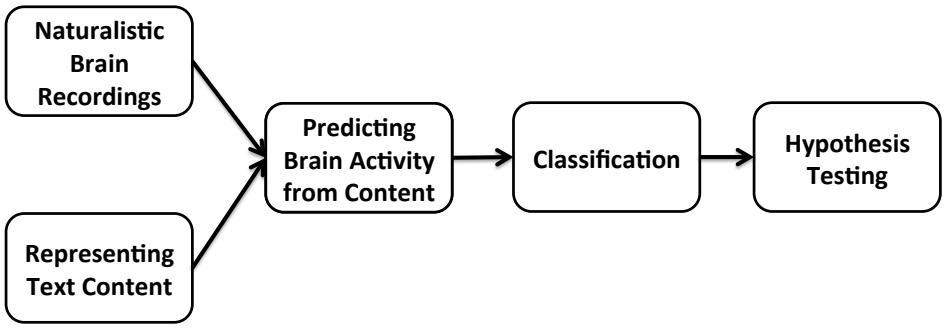
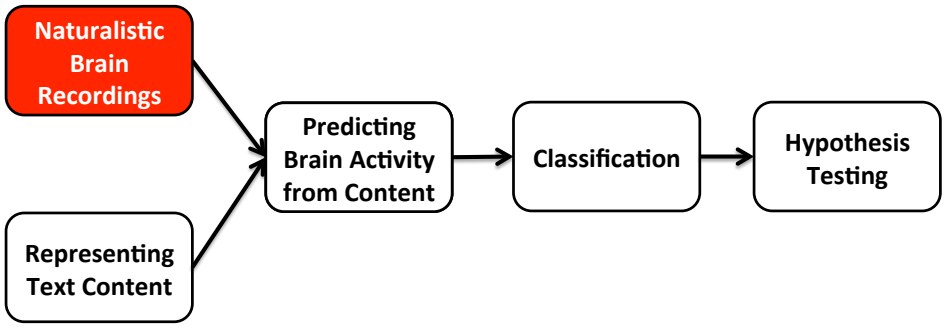


Figure 3.1: Diagram of the main steps of the experimental pipeline.

In this section we describe this generative modeling approach. We begin by the description of the experimental setup for naturalistic data collection. We then explain how we model the content of the naturalistic stimulus in a space that allows for generalization of our predictions to unseen stimuli. Next, we move on to the data predicting step in which we express the brain responses as a function of the input stimulus. We test out predictions with classification tasks, and we finally use those tasks to infer conclusions about brain representations using various hypothesis testing strategies.

3.1 Naturalistic experiment design: trading off repetitions for richness of stimulus



Traditional functional neuroimaging studies typically consist of highly controlled experiments which vary along a few conditions. The stimuli for these conditions are artificially designed, and therefore might result in conclusions that are not generalizable to how the brain works in real life. When studying language processing for example, very few experiments show subjects a real text, and show instead carefully designed stimuli.

Furthermore, the analysis of functional neuroimaging data has typically consisted in simple comparisons: regions which respond differently to the individual conditions are identified. Many researchers have recently started using *brain decoding* (i.e. classifying the stimulus being processed from the subject's brain image), which can reveal responses encoded in subtle patterns of activity across a brain region. However, brain decoding is still mostly used in a rather limited fashion. In order to predict which condition an image corresponds to, a classifier is trained on several examples of each condition. This classifier is *not* able to generalize its knowledge to *novel* conditions not seen in training. It can therefore be argued that such a model does not represent a broad understanding of brain function.

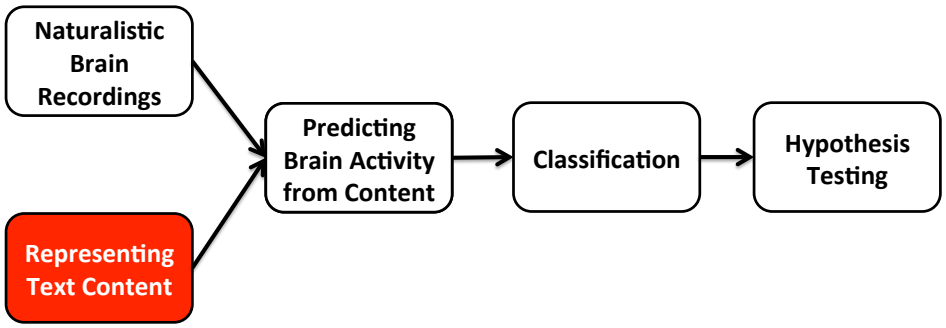
Instead of highly controlled experiments, naturalistic design aims to reproduce the natural conditions under which the brain processes the task of interest. Studying reading in an "ecologically valid" setting (i.e. when subject read real text) would hopefully reveal more insights about how the relevant brain processes. However, it presents with a fair share of difficulties. The imaging tools might be too slow for the dynamic process of reading at a natural pace and unable to identify the contributions of individual words or concepts to brain activity (e.g. fMRI acquires one image every 2 seconds, and is measuring a delayed smooth hemodynamic response). The imaging tools might also be very noisy and necessitate multiple repetitions of a stimulus to procure a reliable image, and repetitions makes the stimulus less diverse and natural.

Another big difficulty is that uncontrolled experiments are, well, uncontrolled. While it is true that a natural text will consist of a rich sample of linguistic properties, these different properties will be correlated because (a) the experimental text cannot be very long because the data acquisition time is limited to a couple hours and (b) these linguistic properties could be inherently correlated (e.g. the part of speech of a word and its semantic properties). Effectively controlling for the other variables in a classical controlled experiment allows the researcher to clearly identify the contribution of the variable of interest to the brain activity. However, it could be

argued that it is hard to create language stimulus that varies along only one dimension (e.g. text with more complicated syntax but exactly the same meaning). Furthermore, it might be that the correlation of these variables is inherent to language and should be accounted for.

The experiments in this thesis study natural text. Chapters from popular fiction are presented to subjects in an fMRI or MEG scanner. Brain activity recordings are continuously acquired while the subjects read the text one word at a time. Words are presented at a rate that is close to the natural reading pace. This is in stark contrast with most other experiments: the usual experimental practice is to have defined stimuli of a few or several seconds and repeat each of them multiple times [79, 117]. This is done because of the low signal to noise ratio of neuroimaging data: the repetitions are averaged to obtain a brain image that is less noisy. In our experiments however, the stimulus is presented to each subject only once. This allows us to present a long and rich passage of text that is less likely to be biased on a given dimension (e.g. word frequency, sentence length etc). The hope is that the meanings and concepts of interest will occur sufficiently frequently in the natural text that it would be possible to detect their individual contribution to brain activity from the noisy signal.

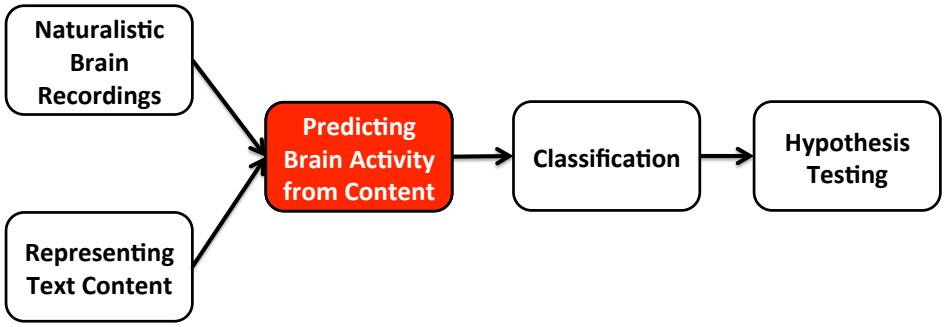
3.2 Intermediate feature space enabling zero shot learning



A primary characteristic of our approach is that it is generative. We show the subjects a complex stimulus, with each part occurring only once. From a subset of the data, we are able to learn a model that can be generalized to unseen experimental text. Our model predicts the brain activity related to reading a passage as a function of the content of that passage. Therefore, an expressive annotation of the experimental text is necessary. This annotation will constitute an IFS. For example, we can label all the words in the text with their part of speech (e.g. noun, verb, adverb) and their grammatical role in the sentence (e.g. subject, object, noun modifier). These labels will constitute a syntactic IFS, we can use it to learn the brain responses associated with different parts of speech or grammatical roles. Since other texts can also be expressed in this same syntactic IFS, we can now predict the brain activity for these other texts as a function of the responses associated with different grammatical features. As we mentioned previously, this method is conceptually different from a simple decoder. If the decoder is a simple classifier that learns to differentiate between a small number of stimuli based on several examples of each class, then this decoder will be able to classify unseen examples of those classes, however, it will not be able to generalize to new concepts.

We gave above an example of a syntactic IFS, but reading a natural text involves much more than understanding its syntax. We will cover in this thesis many other IFS, covering the semantic and narrative properties of the text, as well as IFS that try to separately account for the properties of the word and its context. However, all these investigations will rely on the expressive ability of IFS modeling to generalize to novel stimuli. This method has been used to study the representation of words in [79] and [94], where it is knows as zero-shot learning. This name refers to the fact that a classifier is able to guess the word associated with a brain image without ever seeing it in training. We will explain these steps in detail in the remainder of this chapter.

3.3 Predicting brain activity associated with textual content



In this section, we focus on the predictive model that allows us to express brain activity as a function of the content of the stimulus. We will consider fMRI data which consists of 3D images of tens of thousand of **voxels** (volume pixels) acquired at a rate of one or two seconds (this time is called **TR**, or time to repetition). We will also consider MEG data which is acquired at different **sensors** on the surface of the head. MEG data is typically acquired at a fast rate (e.g. 1KHz). MEG data can also be localized to **sources** of activity in the brain of the subjects. These sources are inferred based on the anatomical scans of the brains of the subjects using **Minimum Norm Estimation (MNE)** [48].

Predicting trial based fMRI activity

We introduce notation consistent throughout the thesis and note that we refer to real valued variables by lower case letters without boldface, vectors as boldfaced lower-case letters and matrices in boldfaced upper-case.

We use here the term **trial based experiment** to refer to experiments with clearly defined trials such that we can obtain an independent image for each stimulus presentation. In Mitchell et al. [79] such an experiment is used: brain activity is collected as native English speakers as look at word-picture combinations, specifically sixty concrete nouns (e.g., “apple”, “car”), accompanied by black-and-white line drawings of those objects.

FMRI measures the hemodynamic response, a change in the blood flow in a region of the brain due to neural activity. This change is slow and lasts about 10 seconds, peaking about 5 seconds after the stimulus is presented. In [79], the latency of the hemodynamic response

was handled by averaging the activity acquired 4–8 seconds after stimulus onset, resulting in a single brain image per subject per stimulus per exposure. For each of the different objects in the experiment, a semantic vector is constructed using a large corpus of text. This vector is 25 dimensional, each entry corresponds to the frequency with which each object occurs next to one of 25 verbs in the corpus. The verbs approximate the space of interactions that are possible with those objects (eating, holding ...). These vectors therefore constitute a semantic IFS.

For a trial based experiment such as the one above, we assume a linear model: the average hemodynamic response y_{vt} of voxel v to the stimulus displayed at time t is a linear combination of that stimulus's features denoted by the P-dimensional feature vector $\mathbf{x}_t$,

$$y_{vt} = \mathbf{x}_t^\top \boldsymbol{\beta}_v + \epsilon_{vt},$$

where $\boldsymbol{\beta}_v$ is the P-dimensional regression coefficient vector of v and ϵ_{vt} is mean-zero noise for voxel v at time t, with variance σ_v^2, combining measurement error corrupting our observation with fluctuations and the effects of specification error. Finally, we assume that the ϵ_{vt} has a Gaussian distribution. More succinctly, we will stack the $\mathbf{x}_t$s into a $T \times P$ matrix $\mathbf{X}$, and for each voxel v, write its activity over the course of the experiment as a T-dimensional vector $\boldsymbol{y}_v$.

Predicting continuous fMRI activity

For an experiment with continuous recording, such as the natural reading experiments we are interested in, words are presented at short intervals compared to the latency of the hemodynamic response and the sampling frequency. Therefore their individual contributions to the fMRI signal are superposed. We will model therefore the activity y_{vt} of voxel v at time t as a linear function of the recent history of the stimulus. We call such an experiment in which the stimulus cannot be neatly divided into trials a **continuous fMRI** experiment.

We aim to find the mapping between the different types of IFS (e.g. semantic, syntactic, narrative) and the neural activity y_{vt}. We want to learn the response of this voxel v to a given IFS vector $\mathbf{x}_t$. An IFS has P dimensions. Consider an individual dimension j, corresponding to the j's text feature (e.g. if we are dealing with a semantic IFS, the j's feature can be edibility). We first assume that the text feature j has a signature activity in voxel v that is consistently repeated every time the brain encounters this feature (for the regions that do not encode this feature, we will ideally learn a signature activity equal to 0). Due to the TR = 2 seconds we use in our natural reading experiments, and the typical latency of the hemodynamic response, we are only interested in the points of the response signature that are sampled 2, 4, 6 and 8 seconds after the onset of feature j (β_1^{vj}, β_2^{vj}, β_3^{vj} and β_4^{vj}). See Fig. 3.2(a). It is important to note that we do not constrain the shape of the learned response signature. We also tried estimating the response with 5 time points (2 to 10 seconds after onset) and 6 time points (2 to 12 seconds). However this manipulation did not significantly change the performance and therefore we use 4 time points for computational and statistical reasons[1].

[1] This experiment utilizes the complex pipeline explained in this chapter, and utilized in the next chapter. After computing the accuracy and the chance distribution at every voxel and for every feature, we repeat the entire experiment with more estimated time points per response signature: 5 and 6, corresponding respectively to points 2 to 10s and 2 to 12s after feature presentation. While the obtained patterns of representation vary slightly, we do not find

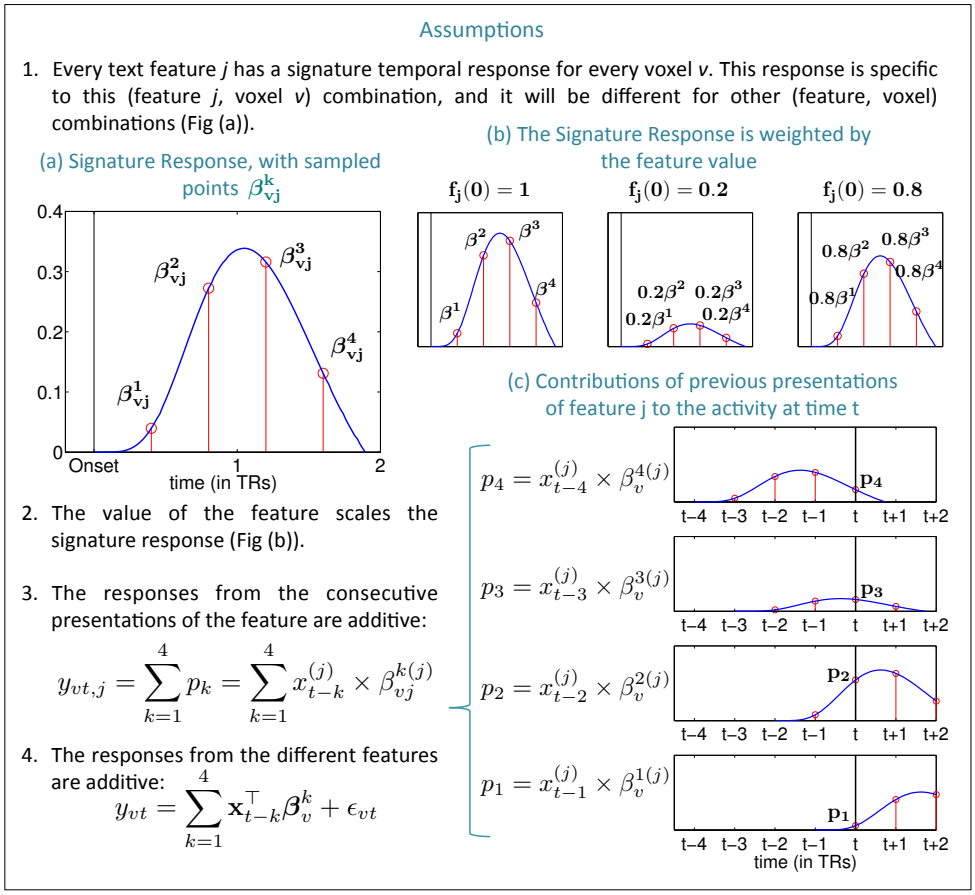


Figure 3.2: **Time model of a voxel's response to the consecutive occurrences of the features of a story**. Because of the hemodynamic response latency, the occurrence of a feature at time t will affect the activity of the voxel for several TRs after time t. This latency is accounted for by considering occurrences of features at previous TRs when modeling a voxel's activity at time t. (One TR is the repetition time needed to acquire one fMRI image, here we use a TR of 2s).

The second assumption is that the signature activity is scaled by the value of feature j at the time the feature is presented. See Fig. 3.2(b). Moreover, we assume that the responses created by successive occurrences of a feature are additive. The contribution of text feature j to the activity at time t in voxel v is:

$$\sum_{k=1}^{h} x^{(j)}_{t-k} \times \beta^{k(j)}_v, \tag{3.1}$$

where $x^{(j)}_t$ is the value of feature j at time t. Another way to think about this is that the activity created by the feature is the convolution of the response signature with the time course of the

any region in which there is a significant improvement for using either type of window. Since the performance is not different, we chose to use 4 because of statistical concerns: we have a training set of about 1100 points and 195 features, it is more advisable to limit the amount of covariates when estimating the model.

feature. Above we considered the brain activity to be created by one story feature. Now we include the activities created by all of the features we have defined above, again assuming they are additive. This gives the model:

$$y_{vt} = \sum_{k=1}^{h} \mathbf{x}_{t-k}^{\top} \boldsymbol{\beta}_{v,k} + \epsilon_{vt}. \tag{3.2}$$

We therefore model the voxel's activity $y_v(t)$ as a linear combination of the values of all the features at times $t-4$ to $t-1$. We know the time courses of the feature values and the voxel's activity, and we need to predict the set of response signatures.

Our approach is similar to Hidden Process Models [60] that also use a multiple regression setup. The neural activity there is also assumed to be generated by linearly additive processes and all instantiations of the same process share the same response, but unlike the case of our model, the delay in the onset of the response is variable.

Our model can be put in a form more similar to the static case by regressing $\boldsymbol{y}_v$ on the vector obtained by concatenating $\mathbf{x}_{t-1}, \mathbf{x}_{t-2}, \mathbf{x}_{t-3}, \mathbf{x}_{t-4}$ into a single P'-dimensional feature vector $\bar{\mathbf{x}}_t$. We can similarly concatenate the regression coefficients for this concatenated feature vector to get a P'-dimensional regression vector $\bar{\boldsymbol{\beta}}_v$. We overload notation to refer to $\bar{\mathbf{x}}_t$ and $\bar{\boldsymbol{\beta}}_v$ as $\mathbf{x}_t$ and $\boldsymbol{\beta}_v$, since from this point the methods apply to both a trial-based approach and a continuous modeling approach.

Predicting MEG activity

MEG records the change in the magnetic field on the surface of the head as a result of brain activity. This is an instantaneous recording, and therefore there is no need to account for the latency of the measurement. However, the sampling frequency is now greater than the stimulus presentation rate. This means that for every stimulus t, we will have measurements with different lags a (e.g. if we present a word every 500ms, we will have 500 recordings for each word t, one every 1ms). We can use a similar model to the trial based fMRI case:

$$y_{vat} = \mathbf{x}_t^{\top} \boldsymbol{\beta}_{va} + \epsilon_{vat},$$

where y_{vat} is the activity at sensor (or source) v for stimulus t, at time lag a.

We might however want to account for the contribution of previous stimuli to the activity recorded for stimulus t at time lag a (for example the brain might still be processing previous words). In that case, we can add the feature vectors of the previous words $\mathbf{x}_{t-k}$ for an appropriate set $k \in \mathbf{K}$ when constructing the $\mathbf{X}$ matrix, similar to the continuous fMRI case.

Learning the models

In equation 3.2, we did not consider different subjects, and only considered a hypothetical voxel v. However, in reality, we have S subjects, and $V_T^{(s)}$ voxels for each subject. The regression in equation 3.2 can therefore be rewritten as:

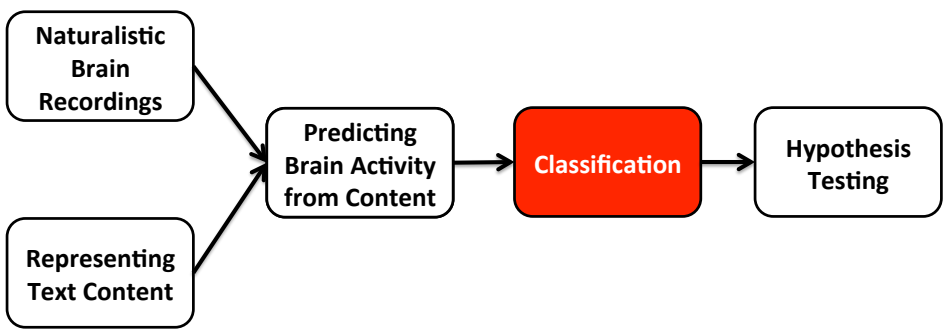
$$\mathbf{y}_v^{(s)} = \mathbf{X} \times \boldsymbol{\beta}_v^{(s)} + \epsilon_v^{(s)} \tag{3.3}$$

where:

- s is the index of a given subject ($1 \leq s \leq S$)
- n is the number of TRs (or time points)
- $\mathbf{y}_v^{(s)}$ is the $n \times 1$ vector of activity of voxel v of subject s
- $\mathbf{X}$ is the $n \times K$ matrix of text features (these could either correspond to the features of the current stimulus or of the history of the stimulus, for example in the continuous fMRI case, every row contains the features of the 4 previous TR, i.e. $K = 4 \times F + 1$, for the intercept term)
- $\boldsymbol{\beta}_v^{(s)}$ is the $K \times 1$ vector of response signatures in voxel v of subject s
- $\epsilon_v^{(s)} \sim N(0, \sigma_v^2 \mathbf{I}_n)$ is the $n \times 1$ vector of errors (n is the number of TRs) caused by noise in voxel v of subject s (σ_v^2 is the noise variance at voxel v and $\mathbf{I}_n$ is the $n \times n$ identity matrix).

We explore in section 3.5, in detail, different methods for learning the vectors $\boldsymbol{\beta}_v^{(s)}$.

3.4 Testing the predictive model with a classification task



As explained above, matrix $\mathbf{X}$ contains the vectors of the IFS of interest (and might or might not contain the history of the stimulus, depending on the experiment type). The matrix of brain activity data for each subject is denoted $\mathbf{Y}^{(s)}$, and we can concatenate all the matrices into:

$$\mathbf{Y} = [\mathbf{Y}^{(\mathbf{1})}, \mathbf{Y}^{(\mathbf{2})} ... \mathbf{Y}^{(\mathbf{S})}], \tag{3.4}$$

such that $\mathbf{Y}$ contains in **each row t the concatenation of the entire brain images for all subjects, at TR t** (the stimulus is always presented to the subjects with the same timeline). In the case of MEG, the row t will contain all the images recorded while word t was on the screen, i.e. as many images as time lags. Similarly:

$$\mathbf{B} = [\mathbf{B}^{(\mathbf{1})}, \mathbf{B}^{(\mathbf{2})} ... \mathbf{B}^{(\mathbf{S})}], \tag{3.5}$$

where

$$\mathbf{B}^{(\mathbf{s})} = [\boldsymbol{\beta}_1^{(s)}, \boldsymbol{\beta}_2^{(s)}, ..., \boldsymbol{\beta}_{V_T^s}^{(s)}] \tag{3.6}$$

3.4.1 Cross-Validation Procedure

To test the validity of the learned response signatures, we constructed a binary classifier that decodes the text being read from a given brain data frame. We start by partitioning the timeline into 10 cross-validation folds. Then for every fold i:

1. If the data is acquired in a continuous fMRI fashion, separate the test (fold i) and training data (other 9 folds) by discarding the data corresponding to 5 TRs before and after fold i.
2. Use the training data to estimate the response signatures of all features in all voxels and all subjects ($\mathbf{B}$), using one of the methods in section 3.5. **It is important to note that the responses are learned independently for each voxel (except for some of the later methods) and each subject**. Also note that the penalty parameter for each voxel that is described for any of the methods in section 3.5 is chosen using only the training data.
3. Divide the timeline of fold i into non-overlapping time windows, each of length L time points. Then, for every pair of L-long segments:
   (a) Take the two test text-frames ($\mathbf{S_1}$ and $\mathbf{S_2}$) and predict the corresponding brain activity using the learned responses $\mathbf{B}$, as shown in Fig. 3.3.
   (b) Use the two predictions $\mathbf{P_1}$ and $\mathbf{P_2}$ to classify each of the two test data-frames $\mathbf{T_1}$ and $\mathbf{T_2}$ independently: i.e. assign to each data-frame the text-frame with the closest prediction, using a distance function explained in the following subsection.

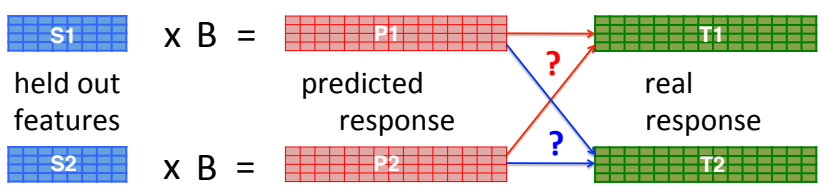


Figure 3.3: Diagram of the classification task. The task is to assign to each held-out data segment of length L ($\mathbf{T_1}$ and $\mathbf{T_2}$) the $L \times R$ seconds portion of the text to which it corresponds (one of the the two dark blue segments, R is the sampling rate). This is done by predicting the activity using the learned weights, then computing the distance between the two predicted responses ($\mathbf{P_1}$ and $\mathbf{P_2}$) and the real segment. The classification of $\mathbf{T_1}$ and $\mathbf{T_2}$ is done independently, i.e. for $\mathbf{T_1}$, the story passage S_1 or S_2 is chosen, and then, in a different test, for $\mathbf{T_2}$, the story passage S_1 or S_2 is chosen.

We average the results of all the cross-validation folds and obtain an overall classification accuracy.

Note: choosing the blocks of length L to be continuous is was made because of the continuous design. In a trial based design one could pick a random set of L images, matched with a random set of another L images.

3.4.2 Classification Procedure

Here we describe how the distances between a test segment $\mathbf{T}$ and the two predicted segments $\mathbf{P_1}$ and $\mathbf{P_2}$ that we compare it to are computed (see Fig. 7). We discuss the methods here in terms

of voxels. However, these can easily be adapted to MEG sensors (or sources) as we explain in chapter 5. We use two methods:

- **Whole-Brain** classification:

  The simplest way to perform classification is to use all the voxels from all the subjects in order to determine the distance between the predicted segments and the true segment. Because we are working with single trial data, concatenating the voxels from different subjects in a row acts as a substitute for multiple repetitions. We compute the Euclidean distance between the two images: $||\mathbf{T} - \mathbf{P_1}||_\mathbf{2}$ and $||\mathbf{T} - \mathbf{P_2}||_\mathbf{2}$.

  Importantly, this test method combines data from multiple subjects without averaging data over subjects in either the learning step (as we saw above) or the classification step. The multi-TR segment that we are classifying is actually a multi-TR concatenation of brain images from all subjects, instead of a multi-TR segment of one brain's images. Since every voxel in that data is trained independently, and contributes to the Euclidean distance independently, then this concatenation does not make any assumptions on the subject's alignment.

  We might want to boost the accuracy when using smoothed data by voxel selection, in the following way. At every cross-validation fold, we use the training data in order to find the best subset of voxels to use. This is done via a nested cross-validation step on the training data what determines which voxels have the best accuracy and how many of the top voxels to use to obtain the best combined accuracy. These voxels are then used to classify the untouched test data: we compute the Euclidean distance using only the columns that correspond to these voxels.

- **Concatenated Searchlight** classification:

  Whole-Brain accuracies do not tell us about which parts of the brain are contributing to the classification accuracy. In order to assess this, we perform the classification "locally", looking in one region of the brain at a time. Regions are defined as $k \times k \times k$-voxel cubes centered around one MNI voxel location, k being an odd integer. This method is similar to the Searchlight approach commonly used in neuroimaging [65], however we expand it to include data from multiple subjects (and in chapter 5 we will modify it for use with MEG):

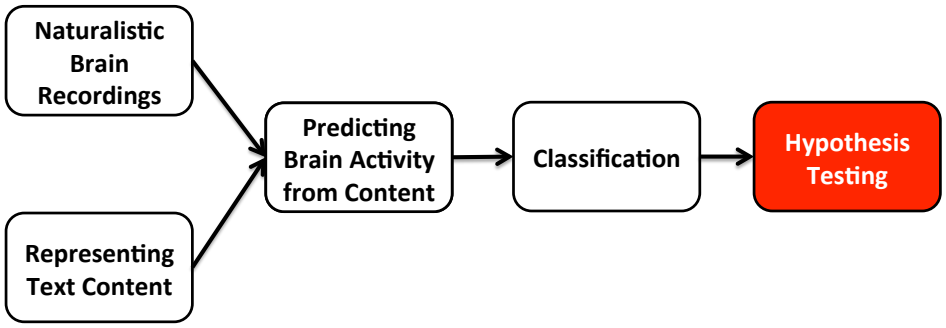
  - We pick a cube size k: for example, $k = 5$ gives a $5 \times 5 \times 5$ voxels cube (to look at one voxel at a time we take a $1 \times 1 \times 1$ voxel cube)
  - For every voxel location (x_i, y_i, z_i), we select the set of voxels whose coordinates fall in the $k \times k \times k$ voxels cube centered around that location. This can be done for each subject independently, in the case where we are interested to look for regions with high accuracy on a single subject basis. It can also be done by selecting the union of voxels from all subjects that fall in this cube. We call the set of voxels selected at this step $\mathbf{V}_i$.

    Because we are working with single trial data, concatenating the corresponding voxels from different subjects in a row acts as a substitute for multiple repetitions. Additionally, since the alignment of the subjects to the same anatomical space is not perfect, taking a $k \times k \times k$ voxel cube with $k > 1$, allows us to circumvent small

variations in the anatomical configuration of the subjects brains.

- For each of these sets $\mathbf{V}_i$ of voxels, we compute the Euclidean distances:
  $||\mathbf{T}(\text{all rows, voxels in } \mathbf{V}_i) - \mathbf{P_1}(\text{all rows, voxels in } \mathbf{V}_i)||_2$ and
  $||\mathbf{T}(\text{all rows, voxels in } \mathbf{V}_i) - \mathbf{P_2}(\text{all rows, voxels in } \mathbf{V}_i)||_2$
  Note: we are performing this computation at every voxel, so we are actually performing N_v classifications, where N_v is the total number of voxels in the experiment. The total number of voxels is the union of all the anatomical locations from all the subjects (since the brains of the subjects might vary in location of their boundaries).

Hypothesis Testing



Once we obtain classification accuracies, we need to perform hypothesis testing. Remember what we are interested in is to find whether a particular process is related to the brain activity at a particular location. We have modeled this process as an IFS and used it to construct a classifier. The main assumption we use now, on which this thesis relies, is that if a region of the brain is not processing a particular language aspect (such as semantics or syntax), then we would not be able to build a good predictive model of brain activity in that region using the relevant IFS. Therefore the classification accuracy would not be significantly higher than chance. If we actually find that the classification accuracy is higher than chance, then this result suggests that the brain region is involved in the process of interest. Because the different text IFS annotations might be correlated to each other or to other variables not accounted for, the higher than chance classification accuracy might be caused by these spurious correlations. We go into methods for avoiding these faulty conclusions in chapter 6.

When dealing with MEG data, we have an additional dimension of interest: the time after stimulus onset. We will be able to extend these hypothesis tests to specific time windows after stimulus onset. For example, we will be able to test a hypothesis such that: the sensors above the temporal lobe are related to processing syntactic features of the words in a text, 300-400ms after the words are presented. We go into detail in chapter 5. In chapter 6, we will also see that such a time-lag dependent analysis is also possible in fMRI, but we will not go into this now for the sake of simplicity.

- **Whole-Brain Classification Accuracy**

To show that Whole-Brain classification accuracy is significantly higher than chance accuracy, which is 50% in this balanced binary classification task, we compute an empirical null distribution. The null distribution that story features cannot predict neural activity is approximated empirically. A common approach to estimate the null distribution is by running a permutation test: the order of the features is permuted before classification and the procedure is repeated a large number of time. This procedure is appropriate in a trial based experimental setting in which the different samples are identically and independently distributed (IID). However, in a continuous experimental setting, the different samples (e.g. different TRs) of our experiment are not IID given that the data is from a time series. The time series of data and of features varies smoothly and therefore the classifier might detect dependencies between them when there is none, because they happen to vary similarly in this finite sample. The commonly used permutation test will not contain such dependencies and therefore will not correct for them, therefore leading to an optimistically biased answer. To solve this problem we use a solution inspired by [13]: we shift the feature time series by N TRs such that $a < N < b$ and compute the classification accuracy. For a and b large enough (e.g. $a = 400$), there will be no real relationship between the time series of data and the time series of features, however the time smoothness will be conserved, leading to better estimates of the variance of chance classification accuracy, which guarantees less false positives.

- **Identifying Brain Regions Correlated with Different IFSs:**

  To find out where in the brain each IFS is useful, we followed a similar training approach as in section 3.3, except that (1) only one IFS (semantic, syntax etc...) was used at a time and (2) we used a concatenated Searchlight procedure at test time with $k = 5$ and using data from all subjects. Precisely, for every voxel location i, we took the cube of $5 \times 5 \times 5$ voxel coordinates centered around that location. We selected the union of voxels from all subjects that have coordinates included in this cube. Therefore, for every location, we performed the classification of 2 segments of size $20 \times |\mathbf{V}_i|$.

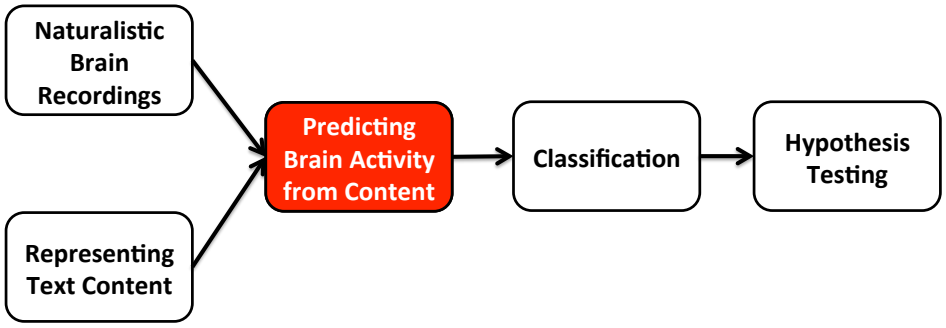
  For every one of these combinations of IFS/subset of data, we obtain a local classification accuracy. We measure significance by computing an empirical null distribution in the same way as for the whole-brain accuracy, then correcting for multiple comparisons using the Benjamini-Hochberg-Yekutieli False Discovery Rate (FDR) [6]. This procedure controls the FDR at level q under arbitrary dependence and therefore we did not need to make independence assumptions about the accuracies of different voxels. The procedure is, for N comparisons:

  - Sort the N p-values.
  - Find the largest j such that

$$p_{(j)} \leq \frac{j}{N} \times \frac{q}{(\sum_{i=1}^{N} 1/i)}.$$

  - Reject the null hypothesis for the j comparisons with the smallest p-values.

3.5 In depth analysis of methods for learning brain responses



We describe in this section multiple methods for estimating brain responses as part of an IFS analysis. This section has been published in [126]. We will use the previously mentioned dataset from [79], which is freely available on the accompanying website of the paper (`http://www.cs.cmu.edu/afs/cs/project/theo-73/www/science2008/data.html`). As mentioned in section 3.3, this experiment scanned native English speakers as they looked at word-picture combinations, specifically sixty concrete nouns (e.g., "apple", "car"), accompanied by black-and-white line drawings of those objects. All nine subjects were exposed six times each to all sixty word-picture stimuli, varying in order. Here the latency of the hemodynamic response was handled by averaging the activity acquired 4–8 seconds after stimulus onset, resulting in a single brain image per subject per stimulus per exposure. The six repetitions of each stimulus are themselves averaged together (within subjects) in the data set.

Each voxel was $3.125\text{mm} \times 3.125\text{mm} \times 6\text{mm}$, and every subject's brain contained $\approx 21,000$ voxels. The subjects' brains were morphed into the same anatomical space, although exact overlap is not achieved due to anatomical differences. One of our methods will utilize the spatial organization of the voxels. For that purpose we divide the voxels into 90 "regions of interest" (ROIs), generally believed to be anatomically and functionally distinct [122]. The ROIs vary greatly in size, from about 20 to about 800 voxels. For ROIs covering a large volume of the brain, the spatial smoothness we hope to exploit is washed out. To counter this, and achieve uniformity of size, we divided ROIs that had more than 200 voxels in half along their largest dimension (x, y or z coordinate). This was repeated as necessary until all regions had 200 voxels or less. After this, we had 191 ROIs.

We used eleven features related to the visual properties of the stimuli (e.g., "amount of white pixels on the screen", "2D aspect ratio"). These annotations were provided to us by the authors of [117], who used the same stimulus set for a different experiment. The original experiment reported these features as ordinal variables on a five-point scale. We selected these features since they represent a fairly coherent set of precisely-measured aspects of the stimuli, ones whose processing is well-understood neurobiologically [112]. For the same reasons, we did not use the many other features also measured in the experiment which are related to semantic or physical properties of the stimuli (e.g. "Is it manmade?", "Can I hold it in one hand?"), as manually rated on the same five-point scale by workers on Amazon's Mechanical Turk crowdsourcing system [117].

In summary, the data consists of sixty words, represented by eleven features each (essentially forming a visual IFS), and their associated average voxel activity across nine subjects.

In previous analyses of this experiments [79], the neural response to reading a word was modeled as a linear combination of the word's features. While such linear models are ubiquitous in fMRI data analyses [4], they have little biological basis. Nevertheless, any smooth model can be locally approximated by a linear regression over a sufficiently small domain, where the range of the feature variables here is fairly small. Plotting actual responses against linear fits shows that the latter are reasonable in these experiments (Figure A.3). Hence, we follow the existing literature in using linear models, and explore multiple ways of fitting and regularizing them — OLS (ordinary least squares), ridge regression, the elastic net, and a hierarchical Bayesian model from small area estimation (SAE). We then consider including the effects of combining these techniques with various forms of spatial smoothing. section 3.5.4 outlines our evaluation criteria for models and their regularizations, by their ability to both predict neural activity from stimuli and to reconstruct stimuli from activity.

Going back to the estimation problem outlined in section 3.3, we need to estimate $\boldsymbol{\beta}_v$:

$$y_{vt} = \mathbf{x}_t^\top \boldsymbol{\beta}_v + \epsilon_{vt}.$$

The residual sum of squares is

$$RSS_v = \sum_{t=1}^{T} (y_{vt} - \mathbf{x}_t^\top \boldsymbol{\beta}_v)^2 = \|\boldsymbol{y}_v - \mathbf{X}\boldsymbol{\beta}_v\|_2^2 \,.$$

where $\|.\|_2^2$ is the squared Euclidean norm. OLS estimates $\boldsymbol{\beta}_v$ by minimizing the in-sample RSS, giving $\hat{\boldsymbol{\beta}}_v = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}_v$. The covariance of the estimates, in a fixed design, is $\sigma_v^2(\mathbf{X}^\top\mathbf{X})^{-1}$.

3.5.1 Ridge regression and Elastic Net

We now review both ridge regression and elastic net, giving the Bayesian counterparts to both. Ridge regression stabilizes OLS estimates via a penalty term [57]. Specifically, the ridge estimator solves

$$\hat{\boldsymbol{\beta}}_v^R = \underset{\boldsymbol{\beta}_v}{\operatorname{argmin}} \; RSS_v + \lambda_v \|\boldsymbol{\beta}_v\|_2^2 \,. \tag{3.7}$$

Equivalently, $\boldsymbol{\beta}_v^R$ is constrained to be small, $\|\boldsymbol{\beta}_v^R\|_2^2 \leq c$, for some $c > 0$. The tuning parameter λ_v controls the degree of regularization. The ridge approach has been used before in neuroimaging with the *same* λ for all voxels [79]. Importantly, in section 3.5.5, we show that tuning λ *separately for each each voxel* improves classification and prediction and provides valuable information about neural organization.

While ridge regression was developed from a frequentist perspective, it has a well known Bayesian interpretation [52]. By imposing a Gaussian prior on $\boldsymbol{\beta}_v$ with prior precision λ, we find

$$\begin{aligned} y_{vt}|\mathbf{x}_t, \boldsymbol{\beta}_v &\overset{ind}{\sim} N(\mathbf{x}_t^\top \boldsymbol{\beta}_v, \sigma_v^2) \\ \boldsymbol{\beta}_v &\overset{iid}{\sim} N(0, 1/\lambda_v \mathbf{I}). \end{aligned} \tag{3.8}$$

Under the formulation in (3.8), the posterior mode coincides exactly with the solution to (3.7). The solution to both formulations has a closed form:

$$\hat{\boldsymbol{\beta}}^R_{v,\lambda_v} = (\mathbf{X}^\top\mathbf{X} + \lambda_v\mathbf{I})^{-1}\mathbf{X}^\top\boldsymbol{y}_v.$$

The covariance is $\sigma_v^2(\mathbf{X}^\top\mathbf{X} + \lambda_v\mathbf{I})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X} + \lambda_v\mathbf{I})^{-1}$, in a fixed-design regression.

The *elastic net* of Zou and Hastie [131] generalizes ridge regression and the lasso of Tibshirani [121]:

$$\hat{\boldsymbol{\beta}}^{EN}_v = \underset{\boldsymbol{\beta}_v}{\operatorname{argmin}}\ RSS_v + \lambda_{1v}\|\boldsymbol{\beta}_v\|_1 + \lambda_{2v}\|\boldsymbol{\beta}_v\|_2^2\ .$$

Setting $\lambda_{1v} = 0$ recovers ridge regression, and $\lambda_{2v} = 0$ recovers the lasso. The L_1 penalty makes $\hat{\boldsymbol{\beta}}^{EN}_v$ sparse, shrinking coefficients on superfluous variables to zero, while the L_2 penalty alone favors small but non-zero coefficients. Again, previous neuroimaging studies favor setting λ_1, λ_2 globally, but we find improved performance by varying them across voxels (section 3.5.5), as chosen by cross-validation (implemented in the `glmnet` MATLAB package by [29]).

As with ridge regression, the elastic net estimate can be viewed as the MAP estimate of a Bayesian model. As shown by Kyung et al. [68], the required prior is a gamma-scale mixture of Gaussians:

$$\begin{aligned} y_{vt} \mid \mu_v, \mathbf{x}_t, \boldsymbol{\beta}_v, \sigma_v^2 &\sim N(\mu_v + \mathbf{x}_t^\top\boldsymbol{\beta}_v, \sigma_v^2) \\ \boldsymbol{\beta}_v \mid \sigma_v^2, \mathbf{D}^*_\tau &\sim N(0, \sigma_v^2\mathbf{D}^*_\tau) \\ \tau_1^2, \ldots, \tau_P^2 &\sim \prod_{j=1}^P \frac{\lambda_1^2}{2} e^{-\lambda_1^2\tau_j^2/2}\ \ d\tau_j^2, \tau_1^2, \ldots, \tau_P^2 > 0, \end{aligned} \tag{3.9}$$

where $\mathbf{D}^*_\tau = \text{Diag}\{(\tau_i^{-2} + \lambda_2)^{-1}\}$ for all i.

3.5.2 Hierarchical Bayesian Small-Area Model

It is biologically plausible that voxels within the same ROI respond similarly to stimuli. Penalization methods, such as the elastic net, make estimates of regression coefficients more precise via stabilization but do not pool information from related voxels. In contrast, techniques for stabilizing parameter estimates by partially pooling information across, or borrowing strength from, related areas have been extensively developed in the literature on small area estimation (SAE; Rao 102). While not traditional in neuroscience, SAE is well known to be effective at shrinkage when there are multiple regions [98], here ROIs. Hence, we explore simple SAE methods for regularization which incorporate ROI-level effects, without completely pooling within ROIs.

The SAE literature typically accomplishes partial pooling using hierarchical Bayesian (HB) models, so we follow that precedent. As before, we model the activity y_{vt} in a voxel v as a linear combination of the stimulus features $\mathbf{x}_t$:

$$\begin{aligned} y_{vt} &= \mathbf{x}_t^\top(\mathbf{z}_v + \mathbf{u}_{A(v)}) + \epsilon_{vt} \\ &= \mathbf{x}_t^\top\boldsymbol{\beta}^{SA}_v + \epsilon_{vt}, \end{aligned} \tag{3.10}$$

where $A(v)$ is the ROI containing voxel v, $\mathbf{u}_a$ is a coefficient vector common to all voxels in area a and $\mathbf{z}_v$ is the coefficient vector specific to voxel v. We have

$$
\begin{aligned}
y_{vt} \mid \boldsymbol{\beta}_v^{SA}, \sigma_v^2 &\sim \mathcal{N}(\mathbf{x}_t^\top \boldsymbol{\beta}_v^{SA}, \sigma_v^2) \\
\boldsymbol{\beta}_v^{SA} &= \mathbf{u}_{A(v)} + \mathbf{z}_v \\
\mathbf{z}_v \mid \nu_v^2 &= \mathcal{N}(0, \nu_v^2 \mathbf{I}) \\
\mathbf{u}_a \mid \alpha_a^2 &\sim \mathcal{N}(0, \alpha_a^2 \mathbf{I}) \\
\sigma_v^2 &\sim \mathcal{IG}(a, b) \\
\alpha_a^2 &\sim \mathcal{IG}(c, d) \\
\nu_v^2 &\sim \mathcal{IG}(e, f),
\end{aligned}
$$

where $a, b, c, d, e,$ and f are user-fixed hyperparameters, and $\mathcal{IG}(\text{shape}, \text{scale})$ is the inverse gamma distribution. Fig. 3.4 shows a plate diagram of the model. The full conditional distributions of all parameters are straightforward (see supplementary materials), so the model can be estimated effectively using partially parallelized Gibbs sampling.

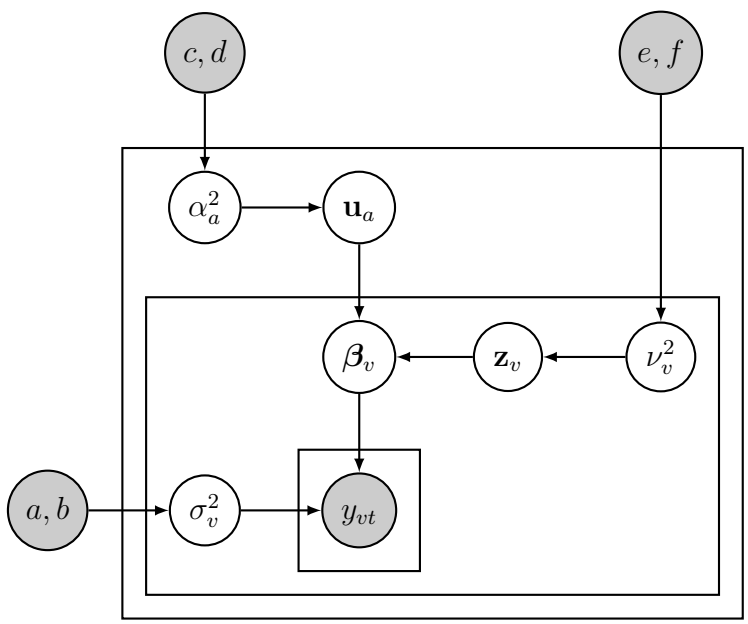


Figure 3.4: Graphical model representation of the small-area model of section 3.5.2.

Just as ridge and the elastic net have Bayesian interpretations, the MAP estimates of this Bayesian SAE model can be seen as a penalized least-squares estimate. Such an estimate is (surprisingly) close to the estimate delivered by ridge regression, for the following reason: The SAE model has a Gaussian prior distribution $\mathbf{z}_v | \nu_v^2 \sim \mathcal{N}(0, \nu_v^2 \mathbf{I})$ for the regression coefficients specific to voxel v, and the voxel-specific variance has an inverse gamma prior distribution, where $\nu_v^2 \sim \mathcal{IG}(e, f)$. Due to this, the *marginal* prior distribution of $\mathbf{z}_v$ is a scaled t distribution, which is well approximated by a Gaussian for reasonable values of the hyper-parameters (see supplementary materials for details). section X of the supplementary materials revisits the statistical

implication of this mathematical approximation, which is that the posterior mode of the HB model must actually be close to the ridge regression estimate.

3.5.3 Spatial smoothing

Neuroimaging data is extremely noisy, and estimates have high variance, even after shrinkage. Much of this noise occurs at high spatial frequencies [4, Chapter 4], and spatial smoothing can help reduce the variance. Since nearby voxels often tend to share activation patterns, spatial averaging may cancel out such noise but maintain signal. Biologically, nearby voxels should tend to respond similarly to stimuli, since recordings of individual cells show that many areas of the brain have a regular spatial organization in their responses to stimuli [112]. While the length scales over which individual neurons' responses vary do not coincide with the sizes of voxels, which general contain many cells with heterogenous properties, it is still the case that nearby voxels should have correlated responses to stimuli. Since the noise in fMRI data is often at much higher spatial frequencies than the signal from voxles, it is reasonable to think that spatially smoothing the activity will enhance the signal-to-noise ratio. This is often done as a preprocessing step [3], but we examine it here as a means of stabilizing parameter estimates.

We explore two kinds of spatial smoothing: *nearest-neighbor voxel-level* and *region-of-interest area-level smoothing*. First we introduce these two forms of smoothing, and then consider smoothed OLS estimates.

Nearest-neighbour voxel-level and ROI area-level smoothing

Nearest-neighbour voxel-level smoothing replaces every voxel by the local average of its nearby voxels. This is done either for the activity levels $\boldsymbol{y}_v$ or the parameter estimates $\boldsymbol{\beta}_v$. Lacking more anatomically-based metrics, we define "nearness" using standard ℓ_p distances of two vectors $\mathbf{r}_1$ and $\mathbf{r}_2$:

$$\|\mathbf{r}_1 - \mathbf{r}_2\|_p \equiv \left(|r_{11} - r_{21}|^p + |r_{12} - r_{22}|^p + |r_{13} - r_{23}|^p\right)^{1/p} .$$

When $p = 2$, this is Euclidean distance and the ℓ_p ball around a voxel contains all other voxels whose centers fall within the given radius. However, when $p = 1$, the ℓ_p ball is a tetrahedral pyramid. We choose a smoothing range or radius separately for each voxel by cross-validation, and replace its value by the average over all voxels within the ℓ_p ball.[2]

Region-of-interest area-level smoothing is defined through solving an optimization problem. Taking the set of regression coefficients in one ROI A, $\mathbf{B}_A := \{\boldsymbol{\beta}_v\}_{v \in A}$, which is a $P \times |A|$ matrix. We penalize large differences between regression coefficients of voxels in the same area. In the Bayesian setting, these are the voxel-wise Bayes estimates. Specifically, for each ROI A, define $\tilde{\mathbf{B}}_A$ as

$$\tilde{\mathbf{B}}_A = \underset{\tilde{\mathbf{B}} = \{\tilde{\boldsymbol{b}}_v\}_{v \in A}}{\operatorname{argmin}} \sum_{v \in A} \|\tilde{\boldsymbol{b}}_v - \boldsymbol{\beta}_v\|_2^2 + \gamma \sum_{i,j \in A} q_{ij}^A \|\tilde{\boldsymbol{b}}_i - \tilde{\boldsymbol{b}}_j\|_2^2,$$

[2]For a given radius, the ℓ_1 ball contains fewer voxels than the ℓ_2, and both are smaller than the ℓ_∞ ball. The latter did so poorly in trials that we only consider ℓ_1 and ℓ_2.

with penalty factor γ and $|A| \times |A|$ similarity matrix $\mathbf{Q}^{\mathbf{A}}$. Fixing $q_{ij}^A = 1$ for all $i, j \in A$, leads to more uniform smoothing. However, letting

$$q_{ij}^A = \exp\{-d(i,j)^2/h^2\}$$

if $i, j \in A$, where d is the Euclidean distance between the locations of voxels i and j and h is a bandwidth, allows closer voxels to be more influential. Since the above optimization problem splits across the dimensions of $\boldsymbol{\beta}_v$, we get P independent optimization problems. Denoting the p-th row of $\tilde{\mathbf{B}}_A$ as $\tilde{\boldsymbol{b}}_p^A$, we find

$$\sum_{i,j \in A} q_{ij}^A (\tilde{\boldsymbol{b}}_{ip} - \tilde{\boldsymbol{b}}_{jp})^2 = \tilde{\boldsymbol{b}}_p^{A\top} \boldsymbol{\Omega}_A \tilde{\boldsymbol{b}}_p^A,$$

where $\boldsymbol{\Omega}_A := 2(D^A - Q^A)$ is twice the graph Laplacian formed using Q^A as the adjacency matrix and D^A as a diagonal matrix whose ith entry is $\sum_j Q_{ij}^A$ [124, Proposition 1]. Hence, $\tilde{\mathbf{B}}_A = (I + \gamma \boldsymbol{\Omega}_A)^{-1} \mathbf{B}_A$. Parameters γ and h are chosen by cross validation.

Smoothed OLS

Since OLS estimates are linear in y_{vt} and covariates are identical across voxels, smoothing $\boldsymbol{\beta}_v$ is equivalent to smoothing y_{vt}. At any voxel v, let $\mathcal{S}_v$ be the set of voxels which are combined with it in smoothing, with the weight of voxel $u \in \mathcal{S}_v$ in the smoothing for v being c_{uv}. These weights are functions of the radius of smoothing in the nearest-neighbor version, or of γ and q for ROI-level smoothing. Then the smoothed estimate at v is

$$\begin{aligned}
\hat{\bar{\boldsymbol{\beta}}}_v &= \sum_{u \in \mathcal{S}_v} c_{uv} \hat{\boldsymbol{\beta}}_u \\
&= \sum_{u \in \mathcal{S}_v} c_{uv} (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{y}_u \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \sum_{u \in \mathcal{S}_v} c_{uv} \boldsymbol{y}_u \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \bar{\boldsymbol{y}}_v \ ,
\end{aligned}$$

which is the OLS estimate with the smoothed response $\bar{y}$.[3]

Despite the simplicity of the technique, smoothed OLS produces results quite comparable to regularization methods such as ridge regression (see section 3.5.5).

The equivalence of smoothing parameter estimates and smoothing the activity does not hold with our other, non-linear estimators. When we report results for combinations of smoothing with other forms of regularization, we are smoothing the parameter estimates.

[3] We are certainly not the first to note that linear smoothing commutes with OLS estimation — see, e.g., Friston et al. [30, p. 12].

3.5.4 Evaluation criteria

Typically, cognitive neuroscientists engage in two forms of predictive inference with fMRI: *forward inference*, of which our IFS analysis is an example, from stimuli to configurations of activity over the brain, and *reverse inference*, from patterns of activity to stimuli. Examples of reverse inference are classical decoding or the indirect decoding we describe in this chapter. While these are often approached as two separate tasks with two distinct sets of models, we perform both forward and reverse inference, using a common model.

Forward inference is a regression problem, where the regression models reviewed above can be applied immediately. Our evaluation criterion for forward inference is the voxel-wise residual sum of squares, normalized by the total sum of squares, i.e. RSS_v/σ_v^2.

Reverse inference is more delicate. As mentioned in section 3.2 we were primarily interested in decoding stimuli from observed neural activity, we could follow the usual practice in fMRI data analysis of estimating "tailored" classifiers or discriminative models [97, 99, 130]. These might be accurate for the particular conditions they were trained on, but by construction they cannot generalize to previously-unseen stimuli, unless they predict as an intermediate step the individual features of the stimuli and then identify the correct stimuli based on the decoded features (Sudre et al. [117]). Moreover, discriminative models do not directly represent anything about how the brain processes information, which is the main point of scientific interest[4]. As shown by [53], the parameters learned in a decoding model, corresponding to each voxel's contribution in a decoding task, cannot be readily used to infer if a voxel is representing a task of interest. For example, some voxels that represent a background process unrelated to the task might receive a high regression weight that serves to subtract that process from the voxels that are informative to the task[5].

We will use the approach described in this chapter to do reverse inference utilizing an intermediate forward inference step. The trained model is faced with the y_{vt} for a held-out stimulus condition in a particular voxel v, and the two sets of features for the correct stimulus condition and another unseen stimulus condition chosen at random. Next, the trained model makes a prediction for both stimuli, and y_{vt} is assigned the stimulus whose predicted activity is closer to the observed y_{vt}. By design, chance performance for the balanced binary reverse-inference task is 50%.

Validation Sets and Cross-Validation We evaluate both forward and reverse inferences with nested 10-fold cross-validation. 10% of the data is held for testing. We then use the remaining training set (90%) to compute the different estimates[6]. If we choose not to smooth the estimates, then we proceed as follows with the training set.

For ridge regression, we use generalized cross-validation [36] to approximate leave-one-out

[4]Symbolically, scientists want to know about $p(Y|X)$, while discriminative models at best give $p(X|Y)$, which, by Bayes's rule, combines $p(Y|X)$ and the distribution of stimuli $p(X)$.

[5][53] do suggest a method to enable a neurophysiological interpretation of the parameters of linear decoding models.

[6]For E2, we throw out 5 images on the boundaries of the training set and the test set to insure that there is no signal leakage from the training to the test set due to the slow decay of hemodynamic responses, causing unintended correlations.

cross-validation error for different λ_v values at each voxel. For the elastic net, we use the ten-fold cross-validation option provided in Friedman et al. [29] to chose the regularization parameters. Finally, for SAE, the level of regularization is determined by the posterior mean variance of $\mathbf{z}_v$, since high variance corresponds to the model being able to chose the parameter freely, i.e. low regularization. The posterior mean variance of $\mathbf{z}_v$ is determined automatically by the Gibbs sampler.

If we choose to smooth the estimates, then in order to pick the smoothing parameter for every voxel and every estimator, we run a nested cross-validation loop. That is, within the 90% training portion of the data, 80% is randomly selected as "inner-fold training" data, and 10% is randomly selected as a validation set. The inner-fold training is done exactly as in the previous paragraph. Smoothing parameters are then set using the average single-voxel classification accuracy on the validation set.

After training, the out-of-sample performance of both unsmoothed and smoothed estimators is reported using the testing set. Thus, the parameters never adapt to the testing set, and we report valid estimates of out-of-sample performance.

3.5.5 Results

Our main findings are as follows: Using cross-validation to pick tuning parameters separately for each voxel,

1. Regularization offers small but real gains in forward prediction;
2. Regularization does not seem to offer improvement in reverse prediction at the individual voxel level, or whole-brain reverse inference;
3. All forms of regularization work about equally well for prediction;
4. Regularization succeeds in making parameter estimates more precise;
5. The spatial pattern of regularization is highly informative: the voxels that are poorly predicted using OLS are the ones that are most heavily regularized under cross-validation.

We explain these points in turn.

Prediction

To summarize, while all models and methods had some predictive ability in our experiments, none of them clearly dominated the others. Model checking, discussed in Appendix A.3, shows that the linearity assumption is reasonable: the actual and the predicted activities tend towards having a linear relationship. Appendix A.3 also shows that the similarity of the results is not due to the models being grossly inappropriate, though they could be somewhat mis-specified. For example, our voxels assignment to ROIs might not be the most optimal one and could have resulted in the small area model not outperforming other methods. Appendix A.3 illustrates this using a simulation in which we show that, if the data is sampled from a small area model, the small area estimates outperform the ridge estimates only when the small area estimator is correctly specified.

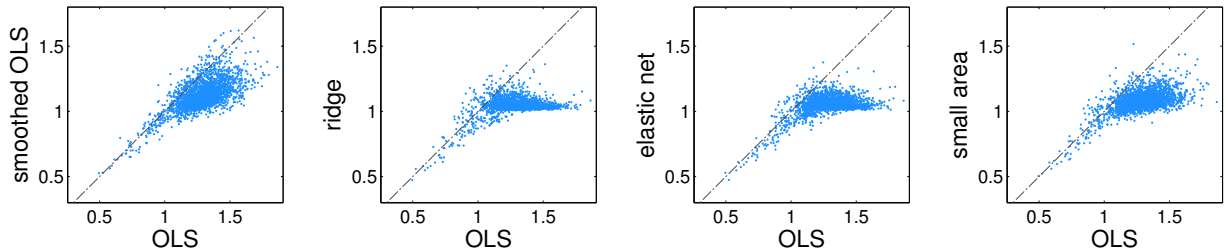


Figure 3.5: Effect of regularization on out-of-sample normalized RSS (RSS/σ^2). Each point represents a single voxel. For each of the plots, the OLS RSS/σ^2 (horizontal axis) is contrasted with the modified RSS/σ^2 after OLS smoothing for ridge, elastic net or small area shrinkage (vertical axis). The four methods result in smaller RSS/σ^2 on average. Furthermore, for all the methods, the predicted activity in the bad voxels (i.e. voxels where RSS/σ^2 is larger than 1) is pushed toward zero. This is visible by the RSS/σ^2 values being reduced toward 1. In other words, shrinkage and smoothing are forcing the estimated parameters to be almost zero if the voxel is noisy and there is nothing that can be predicted.

Forward Inference All our methods had non-trivial ability to do forward prediction in both experiments for some of the voxels, which should be the voxels that are implicated in visual processing (Figure 3.5). All methods of regularizing OLS, including spatial smoothing, led to generally small but significant improvements. The improvement is seen in the noisy voxels: the high RSS in those voxels is greatly reduced when shrinkage or smoothing is used, effectively driving the prediction in the noisy voxels to zero[7]. Combining smoothing with shrinkage did not help forward inference; if anything it often made it worse than either alone (Figure 3.6).

Reverse Inference The effect of regularization on single-voxel reverse inference is ambiguous (see Appendix A.6): accuracy goes up in some voxels and down in others, with no change over all. The classification accuracy of the good voxels varies much less across the different estimators than the accuracy of the bad voxels (see figures in Appendix A.6).

Turning to whole-brain reverse inference, all methods, with and without smoothing, did much better than the chance rate of 50% (Figure 3.7). However, the differences between methods are negligible, and certainly smaller than the fold-to-fold variability of cross-validation. This includes unregularized OLS.

All our methods predict equally well (up to experimental precision), which is surprising. We can rationalize the elastic net performing about as well as ridge regression on the grounds that the former extends the latter by adding an L_1 penalty, which might be unnecessary. Ridge regression is also linked to our hierarchical small-area model via an approximation result (Appendix A.2). However, such connections do not account for why all three forms of shrinkage perform about the same as smoothed OLS or un-smoothed OLS.

We do find a partial explanation from the way we do whole-brain classification (section 3.5.4). Recall that we classify a pattern of activity as belonging to the stimulus whose predicted activity pattern is closest, but weight each voxel in this distance calculation depending on its individual classification accuracy. Thus, the weights are often dominated by a fairly small num-

[7]Since results for both neighborhood- and ROI- based smoothing were nearly identical, we report only those for smoothing over ℓ_2 balls.

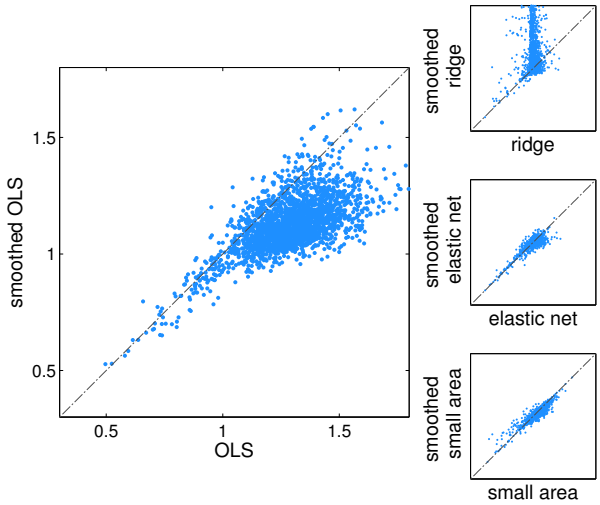


Figure 3.6: Normalized RSS for unsmoothed and smoothed estimators. The larger panels show voxel-wise normalized residuals (RSS/σ^2) for OLS before smoothing (horizontal axis) and after (vertical), showing the value of spatial smoothing for forward inference. The smaller panels consist of the same comparison for ridge regression (top), the elastic net (middle) and the small-area model (bottom), showing that combining smoothing and shrinkage is if anything worse than shrinkage alone. The axes for the smaller panels have been omitted for clarity: they correspond to the larger panels axes.

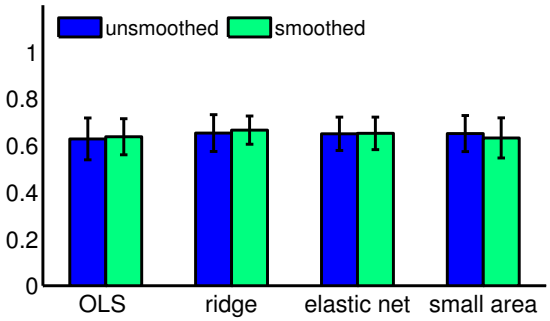


Figure 3.7: Whole-brain classification accuracy, averaging over subjects, for all combinations of estimators and smoothing. Regularization choice or the presence or absence of smoothing don't affect whole-brain classification accuracy.

ber of highly-discriminative voxels. These voxels tend to also be ones where the forward model fits well, and cross-validation or Gibbs sampling selects little or no regularization for them. To support these claims, we examine the effects of regularization on the parameter estimates and the spatial patterns of regularization.

Regularization

Evidence of successful regularization In light of the surprising predictive equivalence of our different methods with each other and with OLS, it is worth verifying that our regularlizers were

in fact regularizing the estimation. From the standpoint of small-area estimation theory, the crucial question is whether the parameter estimates are more precise than the "direct" estimates of OLS. That is, do the new estimates show smaller standard errors, or smaller coefficients of variation, than the direct estimates?

Results like Figure 3.8 are typical across the coefficients and the regularizers. After regularization, most parameter estimates for most voxels had significantly smaller standard errors, sometimes much smaller. This was true even while using cross-validation to pick how much to regularize each voxel. (See Appendix A.4 for additional documentation.)

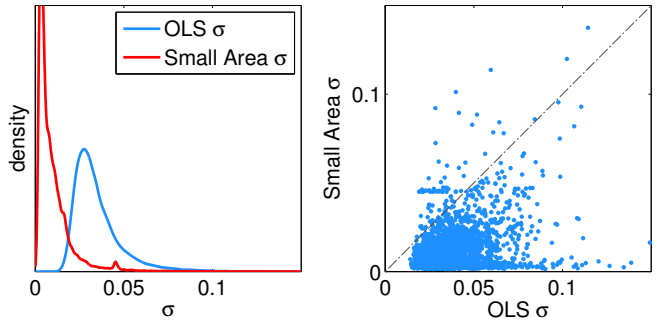
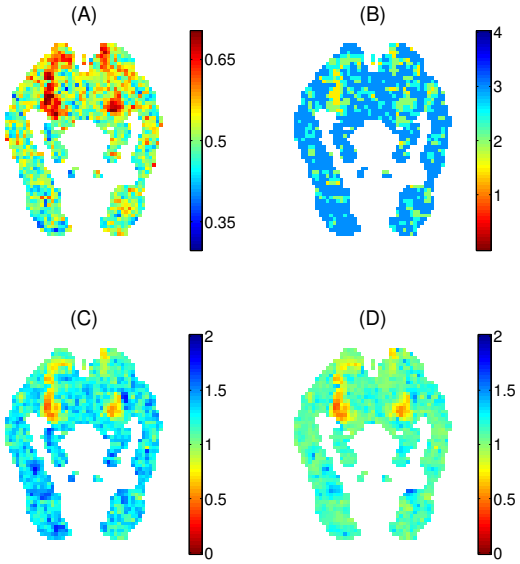


Figure 3.8: Left: Histograms of one regression coefficient's standard errors, aggregating over all voxels, for both OLS and SAE. The sharp peaking of the SAE histogram, to the left of the OLS histogram, indicates that the typical parameter estimate has been made much more precise by the hierarchical model, since the standard errors are much smaller than for OLS. Right: scatter-plot of the same standard errors for OLS and the SEA, this time plotted for each voxel separately. Most of the points fall below the diagonal, so most parameters are being estimated more precisely. Other coefficients and methods of regularization behaved similarly.

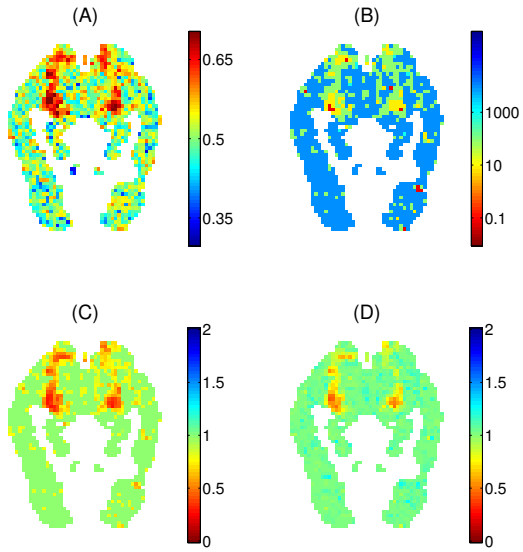
Spatial patterns of regularization and their implications The strength of regularization chosen by cross-validation is not uniform or even random across the brain. It shows quite pronounced, and informative, spatial structure, closely connected to how well voxels predict without regularization.

Fig. 3.9 depicts the relationship between the degree of regularization imposed by our methods, and several measures of predictive accuracy. For two horizontal slices of the brain, these figures illustrate how classification accuracy varies, how strongly regularized each voxel is, and how well the regression model does in and out of sample. (The accuracy plot is omitted for the elastic net to show both penalty factors.) We show in the supplementary material the corresponding plots for the entire brain. The plots provided are for one subject. The other subjects present a very similar pattern of correspondence between voxels with high performance and weak regularization.

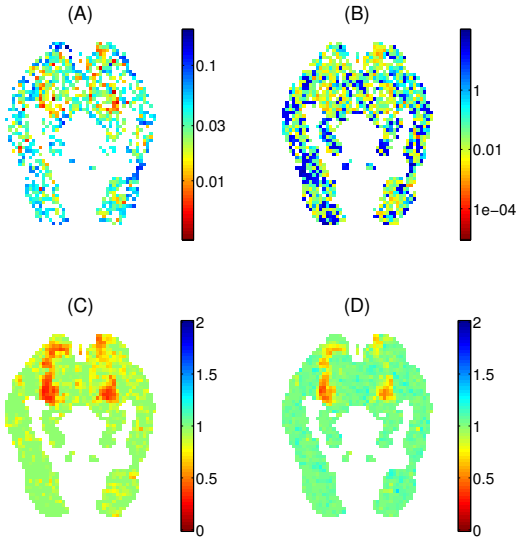
As Figures 3.9a–3.9d show, there is an inverse relationship between predictive performance (sub-figures A and D) and the degree of regularization (sub-figures B) that was chosen by cross-validation. Thus, voxels with stronger signals (as reflected by higher accuracy) needed less regularization. Voxels with high accuracy (3.9a, 3.9b and 3.9d, part A) and especially voxels



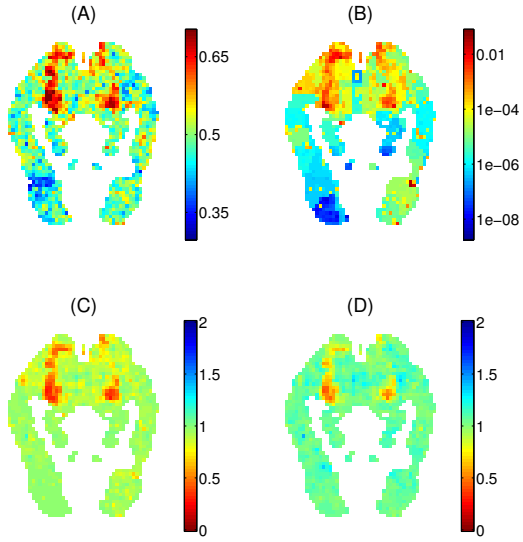
(a) **OLS:** A, classification accuracy; B, smoothing radius; C, D, normalized out-of-sample RSS pre- and post- smoothing



(b) **Ridge:** A, classification accuracy; B, λ parameter; C, D, normalized RSS in- and out-of- sample



(c) **Elastic Net:** A, λ_1 (lasso penalty); B, λ_2 (ridge penalty); C, D, normalized RSS in- and out-of- sample



(d) **Small Area:** A, classification accuracy; B, posterior mean variance of $\mathbf{z}_v$; C, D, normalized RSS in- and out-of- sample

Figure 3.9: Voxel-wise results for each method along one horizontal brain slice. Color schemes are flipped so that red always represents "good" and blue, "bad". Note the similar patterns of classification accuracy in plots a-A, b-A and d-A. Also note how predictive performance (sub-figures A and D) is inversely related to the degree of regularization in every case, whether that is the smoothing radius for OLS (a), the λ penalty for ridge (b), the λ_1 and λ_2 penalties for the elastic net (c), or the small area model (d), where low regularization corresponds to a high variance parameter, i.e., good voxels are allowed to pick their parameters freely. (For the elastic net, good voxels have more lasso-like penalties, as they are voxels sensitive to *some* of the stimulus features.) Smoothing acts as a regularizer for OLS, as seen by the reduced prediction in the bad voxels from subfigure a-C to subfigure a-D. Finally, see that in many cases the in- and out-of- sample errors for "good" voxels are nearly the same.

with low prediction error (sub-figures D) are sparse and spatially clustered. Other voxels are by comparison noisy and more heavily regularized (sub-figures B).

The correspondence between good classification accuracy and weak regularization explains the single voxel accuracy results mentioned in section 3.5.5 and in appendix A.6. In good voxels, classification accuracy is not significantly affected by regularization since the penalty parameter is weak. In the bad voxels, the strong regularization forces the model to learn near-zero weights, and the leftover noise has "random" effect on the single voxel classification accuracy, some times resulting in slight improvement, and sometimes in slight decrease.

For all the subjects, the predictive voxels are clustered in the occipital cortex, which is well-known to be heavily involved in visual processing [112].

Conclusions: Ridge and SAE

We have shown that different forms of regularization predict about equally well. Moreover, they give similar parameter estimates, especially the SAE model of (3.10) and ridge regression. As already explained in section 3.5.2, the marginal prior distribution of β_v is an inverse-gamma variance mixture of Gaussians, which is a t-distribution, where $\beta_v \sim t$. With even a moderate number of degrees of freedom in the t, the marginal prior on β_v is quite close to being Gaussian (Appendix A.2). Similarly, the marginal prior on u_i is also a t distribution. Since β_v and $u_{A(v)}$ are independent *a priori*, the prior on z_v is approximately Gaussian. Since the posterior mode under a Gaussian prior matches ridge regression, the z_v estimated from (3.10) will be close to the ridge regression estimates.[8]

When we simulate from the SAE model, estimating that model shows better forward prediction than OLS or even ridge regression (Appendix A.3.1). The difference between SAE and ridge is small but systematic and significant. However, when the surrogate data from the simulations is re-estimated with erroneous assignments of voxels to ROIs, the advantage of the SAE model over ridge regression vanishes. It *may* be that this is the way in which the SAE is mis-specified, suggesting that a better choice of ROIs would lead to superior prediction. However, we have not been able to rule out other possible mis-specifications.

Conclusions: Computational costs

While our four methods perform very similarly statistically, their computational costs differ by orders of magnitude (Table 3.1). Smoothed OLS and ridge stand out as the most attractive methods, with ridge pulling ahead due to its better behaved out-of-sample residuals.

Our simple and generic HB model is mis-specified, not very firmly grounded in biology, and, as Table 3.1 shows, computationally very costly. With considerable attention to the biology, well-specified models and priors might be crafted for specific applications, though at even greater computational expense. Due to this, we do not advocate the Bayesian approach, unless it could be combined with some way of *quickly* approximating posterior distributions, e.g., variational methods [11, 125] or consensus MCMC [87, 111].

[8] We have not been able to find this approximation result in the literature, but suspect it is a rediscovery.

	cpu time per fold per subject	clock time per fold per subject	total cpu time (with nested CV)
OLS	<1s	<1s	<1min
Ridge	55s	4s	7.5h
Elastic Net	3120s	390s	429h
Small Area	5540s	740s	762h
Smoothing, nested CV	40s	20s	5.5h

Table 3.1: Running times of the various procedures, using 8 Intel Xenon CPU E5-2660 0 cores (at 2.2 GHz), sharing 128GB of RAM. Gibbs sampling for the SAE model was parallelized over the cores.

Consistency of results

While we show here the results on a trial based experiment, these results are very consistent when we analyze naturalistic experiments recorded in a continuous manner. We provide in Appendix A.7 the same results using the naturalistic experiment that we study in chapter 4.

Chapter 4

The Spatial Representation of Language Processes

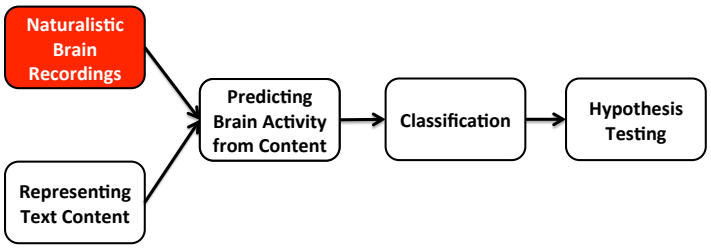
Story understanding involves many perceptual and cognitive subprocesses, from perceiving individual words, to parsing sentences, to understanding the relationships among the story characters. In this chapter, we apply our investigative Intermediate Feature Space (IFS) approach to a naturalistic experiment in which subjects read a chapter from a popular book (*Harry Potter and the Sorcerer's Stone* [106]). We present an integrated computational model of reading that incorporates these and additional subprocesses, simultaneously discovering their fMRI signatures. Our model predicts the fMRI activity associated with reading arbitrary text passages, well enough to distinguish which of two story segments is being read with 74% accuracy.

Reporting accuracy of the trained model predictions is however not the main contribution of this chapter. We also use the brain activity encodings of different story features learned by the trained model – including perceptual, syntactic, semantic, and discourse features – to provide new insights into where and how these different types of information are encoded by brain activity. We align and contrast these results with several previously published studies of syntax, semantics, and models of the mental states and social interactions with others. In this chapter, we use the term "semantic features" to refer to the lexical semantic properties of the stimulus words independent of the context in which they appear, and use "discourse features" to refer to discourse semantics of the story. This work has been published in [127], and we made the data freely available online[1].

Our approach is analogous to [79] that trained a computational model to predict fMRI neural representations of single noun meanings. However, here we extend that approach from single nouns and single fMRI images, to passages of text in a story, and the corresponding time series of brain activity. This work is also analogous to recent work analyzing fMRI from human subjects watching a series of short videos where a large set of objects were identified and annotated with semantic features that were then mapped to brain locations [61], though that work was restricted to semantic features and did not include language stimuli. Our approach is the first to provide a generative, predictive model of the fMRI neural activity associated with language processing involved in comprehending written stories.

[1]Data is available at `http://www.cs.cmu.edu/~fmri/plosone`.

4.1 Experimental design



Material

Participants read chapter 9 of *Harry Potter and the Sorcerer's Stone* [106]. We chose this chapter because it involves many characters and spans multiple locations and scenes. We chose a famous book series because we hypothesized all subjects already had characteristic mental representations of the different characters and locations, and that at least a part of this representation would remain constant throughout the reading of chapter 9. This assumption allows us to use data from the entire chapter to look for the representation of the different characters, e.g. the protagonist Harry Potter. In contrast, had we chosen an unfamiliar story in which we learn about the protagonist's personality throughout the text, the mental representation of this protagonist will arguably change more than Harry's would.

Participants

fMRI data was collected from 9 subjects (5 females and 4 males) recruited through Carnegie Mellon University, aged 18 to 40 years. The participants were all native English speakers and right handed. They were chosen to be familiar with the material: we made sure they had read the Harry Potter books or seen the movie series and were familiar with the characters and the story. All the participants were screened for safety, signed the consent form and were compensated for their participation. Data from one of the subjects was excluded from the analysis because of an artifact that was not removed by our preprocessing procedure.

Design

The words of the story were presented in rapid serial visual format [12]. Words were presented one by one at the center of the screen for 0.5 seconds each (see Fig. 4.1), with punctuation appearing with the word it is collated to. The background was gray and the font was black. We used MATLAB and the Psychophysics Toolbox extensions [9, 62, 96].

The chapter was divided into four runs, of approximately 11 minutes each. Subjects had short breaks between runs. Each run started with a fixation period of 20 seconds in which the subjects stared at a cross in the middle of the screen. The words presentation started after the fixation period. The total length of the runs was 45 minutes, during which about 5200 words were presented. Chapter 9 was presented in its entirety without modifications and each subject read the chapter only once.

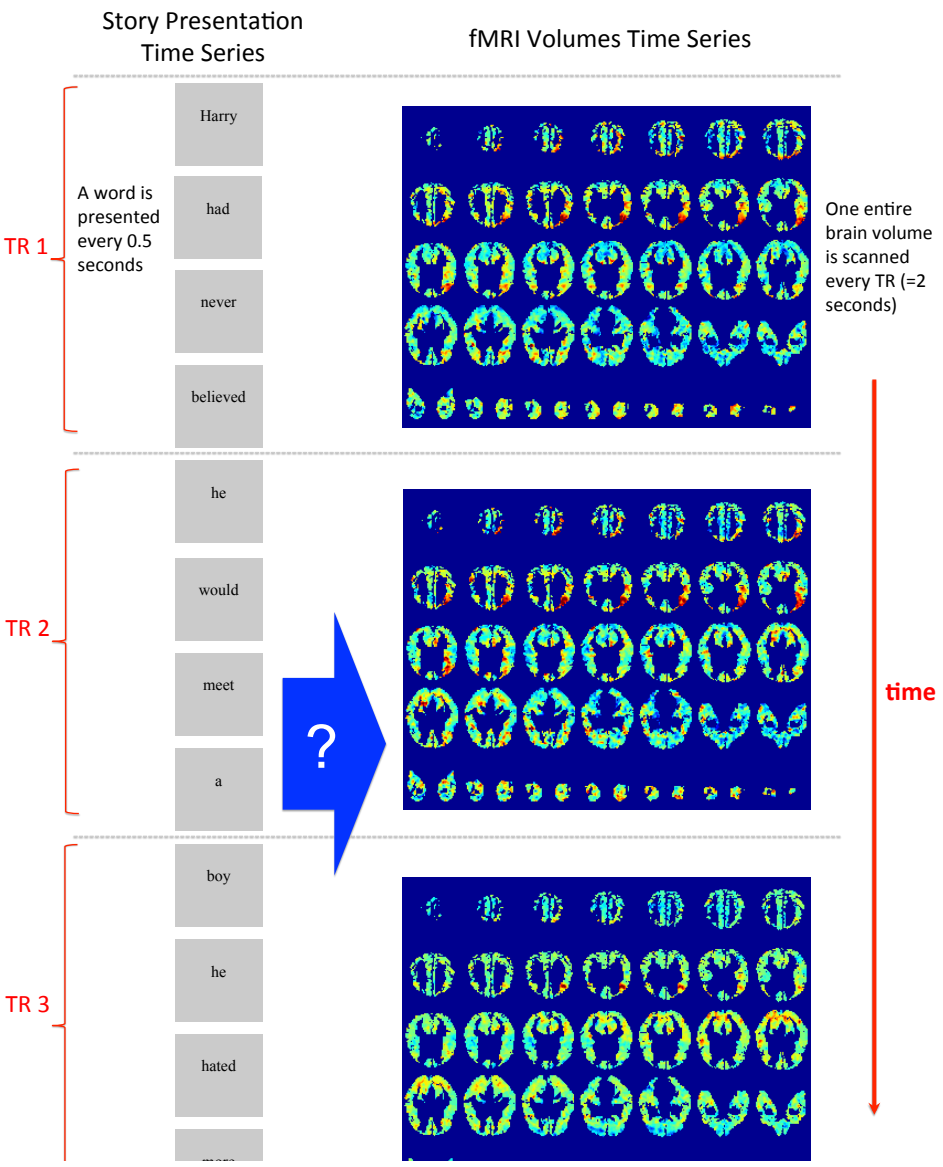


Figure 4.1: Illustration of our fMRI experimental protocol. Words from a story are presented serially for 0.5 seconds each while recording brain activity with fMRI at a rate of one entire brain image each 2 seconds. Our goal is to model how fMRI neural activity during reading reflects the perceptual and conceptual features of the story. Each fMRI activity volume is shown here in 36 horizontal slices. Going right to left through the slices, then bottom-up, corresponds to looking at slices from the bottom of the brain up. Within each slice, the top of the slice corresponds to the posterior of the brain, and the right side of the slice corresponds to the left side of the brain. The images are on a scale from blue to red where blue indicates negative deviation from baseline and red indicates positive deviations. A TR is the time needed to record one brain volume, and is 2 seconds in our experiment.

Before the experiment, we supplied the subjects with a summary of the events preceding chapter 9 and a summary of the main characters and concepts in *Harry Potter and the Sorcerer's Stone* to refresh their memory. We also instructed them to practice rapid serial presentation by viewing a video that replicated the parameters of our design, but with another story (*The Tale of Peter Rabbit* [100]). On the day of the experiment, the subjects were instructed to lay in the scanner and read the chapter as naturally as possible while remaining alert.

fMRI procedure

Functional images were acquired on a Siemens Verio 3.0T scanner (Siemens, Erlangen, Germany) at the Scientific Imaging & Brain Imaging Center at Carnegie Mellon University, using a T2* sensitive echo planar imaging pulse sequence with repetition time (TR)=2s, echo time=29 ms, flip angle=79°, 36 slices and $3 \times 3 \times 3$mm voxels. Anatomical volumes were acquired with a T1-weighted 3D-MPRAGE pulse sequence.

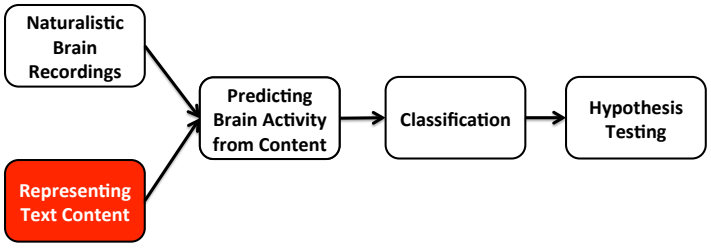
Data preprocessing

We used the MATLAB suite SPM8 [3] to preprocess the data. Each subject's functional data underwent realignment, slice timing correction and co-registration with the subject's anatomical scan, which was segmented into grey and white matter and cerebro-spinal fluid. The subject's scans were normalized to the Montreal Neurological Institute (MNI) space and smoothed with a $6 \times 6 \times 6$mm Gaussian kernel smoother.

Using the Python toolbox PyMVPA [49], we masked the functional data using the segmented anatomical mask, discarding cerebrospinal-fluid voxels. The data was then detrended in MATLAB by running a high-pass filter with a cut-off frequency of 0.005Hz. Visual inspection of the time course of a large number of voxels showed that this threshold was enough to get rid of large block effects and slow trends in the data.

Finally, we selected voxels from each subject, keeping only voxels in 78 cortical Regions Of Interest (ROIs), defined using the AAL brain atlas [122], excluding the cerebellum and white matter. We ended up with an average of 29227 voxels per subject. The anatomical union (number of MNI voxel locations for which at least one subject had a voxel) of these 6 subject's brains was a set of 41073 voxel locations.

4.2 Representing the content of the story



4.2.1 Formal theories of language

In order to build an IFS describing the content of the story, we first turn to how language is formally characterized. The study of language, or linguistics, revolves around the study of the units that language is made of and the rules that determine how they are combined [1]. Linguistics is divided amongst several core areas. Morphology is the study of the structure of words, while phonology studies the structure of sounds in language. Syntax revolves around understanding word categories, phrasal units and sentence structure, and semantics is the study of linguistic meaning.

Other fields of linguistics relate to the fact that language is a dynamic, it's acquired and used by humans for specific purposes. Specifically, the area of pragmatics studies language in relation to the world and how it's used for communication and other functions: it is concerned with how the context of language affects meaning, and extends from studying simple sentences to the study of discourse. Other than pragmatics, linguists study how language is acquired, how it evolves in populations and how it changes across time and populations. However, since the participants in our experiment were all adult, native english speakers, and since the experiment was done at one point in time (it was not a longitudinal study), the evolution of language and language learning are not central to this work and we will omit their description. This subsection is all based on [1].

Morphology and Phonology

Morphology deals with the study of the nature of words and of sub-word structure. Some words are simple and cannot be divided into parts, while others are compound words. Morphemes, the smallest units that can't be subdivided, are either free or bound. Free morphemes constitute a base and can appear by themselves (e.g. "dog", "city", "cold") or as part of another word ("hot-dog", "city-center", "coldness"). Bound morphemes cannot occur by themselves and can be, among others, affixes (e.g.. "un-") or suffixes (e.g.. "-s"). Morphemes (either free or bound) can either be content morphemes expressing a certain meaning (e.g. apple, run) or function morphemes serving a more structural role in the sentence (e.g. "that", "and", "the", "-ful", "re-", "-ed"). The rules for how morphemes are combined depends on many factors, such as the grammatical category of the word (which we will discuss here under "Syntax"). How the meaning of complex words is composed out of the meaning its morphemes is a complex question, since compound words often have meanings that are different than or span more than the set of the meaning of their parts. The meaning of word combinations can be the clear combination of the meaning of its part (e.g. "sidewalk") or different from the original meaning and having a new meaning out of convention (e.g. "hot-dog"), or anywhere in between.

Phonology is the study of the structure and patterns of sounds in language, how to properly describe them, and what is a proper general framework for them. Since we are studying reading and not listening, phonemes, the units of sounds, are not going to be of high interest to us. However, reading is often accompanied by subvocalization, or silent speech, requiring the conversion of the original letter string to phonemes, and therefore modeling the phonetic aspect of the text might contribute to understanding the brain activity related to reading.

Syntax

Syntax is the study of the structural units of language such as the grammatical category of individual words or noun phrases, as well as the rules that govern how they are combined together and the relationships they have. Self-evidently, the structure of sentences is crucial to its meaning, with small variations sometimes changing the meaning completely (consider inverting the subject and object of a sentence, e.g. "the dog bit the man" and "the man bit the dog"), and sometimes, the structure of the sentence is ambiguous because it could be subdivided into subgroups in multiple ways ("The mother of the boy and the girl will arrive soon.", taken from [1], can either refer to one person or two people arriving soon). In order to study sentence structures, linguists have developed methods to construct graph diagrams, most often tree structures, of sentences. These rely on the linear ordering of the words in a sentence, their categorization into parts of speech (whether they are nouns, verbs, adjectives, adverbs etc.) and their grouping into structural constants of the sentence (e.g. in the above sentence, we can have "[The mother of the boy] and [the girl]" or "[The mother of [the boy and the girl]]" as groups).

There is a distinction between the grouping of words into structures, and the grammatical role that these structures have in the sentence, i.e. a noun phrase can serve multiple roles in a sentence, such as the subject of the verb or its object. Identifying the roles of different units in the sentence is therefore a major art of constructing tree diagrams that annotate the structure of sentences. These concepts have been formalized in proper theories, which also deal with issues like structural ambiguity or complex grammatical dependencies that are interleaved in the sentence. Such theories are typically very complex, and we are not trying to select between them this approach. Some computational approaches are more theoretically neutral. For instance, dependency grammar models the structure of sentences by focusing on the link between words in the sentence. Starting with the verb as the root of the sentence, those links summarize the existing relationships between pairs of words and can either be direct, or pass through other words. Dependency parsers (such as the Stanford parser [19] or the MALT parser [92]) are increasingly being used and are are built to handle most cases of structural ambiguities.

Semantics and Pragmatics

The study of semantics, or meaning, has not always had a prominent position in linguistics, perhaps due to the difficulty of stating what exactly it is to "mean". Various theories have been put forth to explain meaning, such as the denotational theory of meaning in which the meaning of a word is the object it refers to. This theory fails to account for the fact that you could use multiple expressions to refer to the same object, and these expressions do not have to have the exact same meaning (e.g. "the first man to walk on the moon" and "Neil Armstrong" refer to the same person but however have different meanings [1]). Another theory sees meaning as the ideas that are in the head of the listener, and has been modified into seeing meaning as the mental images that are evoked, or the concepts that are associated with the word, all suffering from the difficulty of expressing ideas, images and concepts in a manner consistent with how we use the meaning of words, and how we combine the meaning of words into sentences. The sense theory of meaning identifies the "sense" of an expression as different from what it refers to: two expressions can refer to the same object, but their modes of presentation or their cognitive

contents might differ. The use theory of meaning considers the meaning of an expression to be the way in which it is used. There is no real convergence on a theory of meaning, with various theories being more or less incomplete.

Many properties of the meaning of words can be interesting to study in the brain. The brain might help us to understand what happens when a word is polysemous (has multiple meanings) and its meaning can be inferred from the context it falls in. Words have meaning relations: they can be synonymous or antonymous, or they can be included in the meaning of other words, for example, a "dog" is an "animal". When words are grouped together into sentences, their meanings are composed in multiple ways that are not yet accounted for by a complete theory of meaning composition. Sentences can be declarative, interrogative, exclamative or imperative, which brings us to the role that language has when it's being used, and the study of the meaning of language in a context.

Pragmatics studies the function and meaning of language when used in a specific context. For instance, some words like "hello" and "goodbye" are used mainly for the purpose of greeting or bidding farewell, whereas other words might have a literal meaning and another meaning when used as an injunction (e.g the word "brother" when used as "my brother" or in the injunction "oh brother!" [1]). In the message model of communication, a speaker wants to communicate an idea to the listener, he choses a linguistic utterance to express his thoughts and a communication is successful when the listener correctly understands the original idea. The problem with this model is that it fails to explain much of the complexity of linguistic communication. For example, it doesn't explain how disambiguation, non-literal meaning, or speaker intention are accounted for (e.g. when a speaker sarcastically says "That went well!" after a bad turn of events, he means something different than the literal meaning of what he has uttered). The inferential model of communication states that the communication is successful when the listener correctly interprets the speaker's intention. This would rely on a set of shared presumptions and inferential strategies. Amongst others, the listener is presumed to have enough knowledge to infer the intention of the speaker, and the speaker's words are taken to their literal meaning unless there is reason to interpret them otherwise. Indeed, non-literal meaning is a large part of language as evidenced by the frequency of metaphors. Some metaphors are well established and commonly used, while others are novel and created by the speaker to illustrate expressively a concept, and there is interesting fMRI evidence that the brain process these two types differently [73].

We are interested in studying the mental representation of stories. Pragmatics includes the study of discourse, which is any linguistic body that is longer than a few sentences. Sentences in discourse occur in a specific context that contributes to their meaning and to understanding their intent. Discourse can occur in many forms: conversations, letters, speeches, narration, etc. We are interested in studying narrative discourse, which consists in topics sometimes studied in literature, including for example the presence of characters (typically a protagonist with allies and enemies), how the reader identifies with them, of whether the story was told in the first or third person. Narrative discourse revolves around events that usually have a certain chronology and evolve in a certain space, and typically starts with an exposition that is soon interrupted by a conflict or a change, then followed by a climax that is eventually resolved (and sometimes not), providing a certain denouement or conclusion to the story.

4.2.2 Choice of Intermediate Feature Spaces

Now that we have presented the main building parts of language, we can describe and motivate the IFSs that we have used to study story processing. This is the first iteration of an experiment designed to construct a reading map, and therefore, we didn't include all the features that we could have possibly extracted, but mainly the ones that spanned different levels of processing usually studied in the context of reading, as well as the ones that were the most justified, were possible to generate using available NLP tools or by hand and were able to be detected given the experimental design.

We represented our story features as a multivariate discrete time series. We used one TR as a unit of time. This enables us to have the same time scale for the features and the data time series. We compute the value of a feature at any TR by aggregating the features of the four words that were read during that TR (see table 4.2).

The list of all the 195 features we used is provided in table 4.1. This table includes the features that were finally used in the model: a few of our features had too few occurrences and we ended up disregarding them. We also include in table 4.2 as illustration a subset of the feature values for the two segments of the story included in Fig. 4.2(B). Finally, we provide at `http://www.cs.cmu.edu/~fmri/plosone` all these feature annotations (along with the complete fMRI data for our 8 subjects).

Visual features

While designing features for the words in the story, we decided against including morphological and phonological features for several reasons. We didn't have the timing at which each sub-word structure was perceived by the brain since the only measure we have about reading each word is that it's presented for 500ms. We could have used an eye-tracker to measure when each morpheme was being read, however, we know from eye tracking studies of reading that text is usually processed at the level of words or multiple morphemes [105], since readers fixate typically in the space between words. Furthermore, remember that we are using fMRI and acquiring data at a speed of 2s per image, during which 4 words are presented. FMRI is also a slowly varying signal. Even if we could time the perception of different sub-words structures accurately, it would be hard to distinguish their effect on the brain activity. We therefore felt that studying the sub-word structure (morphology and phonology) would be hard in this paradigm that has not been designed with this specific aim in mind. However we did include features aimed at measuring the initial low level processing of words: when reading, the visual signal is first detected by the primary visual cortex, and there is a lot of evidence that the visual input is then processed in the "Visual Word Form Area" [15], making it a key structure in the identification of letters and word forms. To show the implications of these regions in that process, and as a sort of sanity check (we thought these features would be the easiest to detect) we computed simple visual features:

- **Average Word Length:** We compute the average word length in every TR.
- **Word Length Variance:** We compute the variance of word length in every TR.

Syntactic features

Next, we computed syntactic features. As we just saw in section 4.2.1, the structure of a sentence is determined by the structure and grammatical categories of its components, as well as

their grammatical relationships. Typical studies of syntax in the brain typically only use simple measure such as sentence complexity (that can be determined by the complexity of the sentence tree, such as in [10]). We however chose to have rich syntax features that aim to express much more of the structure of the sentence by explicitly describing the structure of words and their grammatical roles.

As mentioned in section 4.2.1, many models of syntactic representation have been proposed that make many different theoretical assumptions, and can include very complex representational structures to encode even simple concepts like "subject". It is hard to distinguish amongst these models and find which of them is closest to the truth. We chose simplicity as a heuristic. To represent word category we used parts of speech, which are simple to find and mostly uncontroversial. For sentence structure, we opted for the use of dependency parsing. This approach is also rather simple and theoretically neutral, is commonly accepted and used in NLP and results in a limited set of possible grammatical roles, making it suitable as an IFS because of its manageable dimensionality.

Using the MALT automated parser [92] we determined the part of speech of every word in the story and obtained the dependency role of every word from the parse tree of the sentences.

We obtained a set of 28 unique parts of speech and 17 unique dependency relationships, for a total of 45 syntactic binary features that indicate if a given part of speech or a dependency relationship occurred within a TR. We also included an additional feature that records the position of a word in the sentence, i.e. its number starting from the beginning of the sentence. This value is averaged for the four words in a TR.

Semantic features

We saw in section 4.2.1 that it is very difficult to describe the meanings of words in an uncontroversial/definitive fashion. Fortunately however, statistical approaches have made it possible to computationally construct vectors encoding those meanings. These distributional semantics approaches rely on the assumption that words with similar meanings are used similarly [107]. An approximation of the meaning of a word can be obtained by the pattern of its occurrence with other words. For example, "apple" is likely to occur with other food items or the verb "eat", but not so likely to occur with building materials or power tools. Word co-occurrences statistics are learned from massive web corpora. These statistics are very large in dimension; one should apply a dimensionality reduction method to make them more manageable.

One one hand, distributional semantics does make some significant simplifying assumptions that might lead to some inappropriate representations (e.g. multiple senses of a word can be merged together). On the other hand however, these methods these methods have a number of advantages. Not only are they of popular use in NLP (where they were show to improve performance in a variety of tasks) and in in computational neurolinguistics [85], but they also make a relatively small number of theoretical assumptions. For instance, there are multiple large lexical databases that are available such as WordNet [78]. These databases have been carefully built by hand. However, because of this, they suffer from requiring a large number of assumptions (and therefore risking various biases), as well as from being incomplete and hard to update.

The distributional semantics approach we use is NNSE (Non-Negative Sparse Embedding) [83], which produces low dimensional representations for word meanings. These representations are enforced to be sparse and non-negative, which makes them more interpretable and cognitively

plausible. The intuition is that, when asked to name the semantic properties of an object, one would list the few salient positive properties (e.g. an apple is a round, usually red, edible object) instead of naming negative properties (e.g. an apple is not a tool), see [83] for more detail.

To obtain these NNSE semantic features, sentences in the corpus are first dependency parsed. Then the co-occurence statistics of words in specific grammatical roles with other words are computed. These co-occurence statistics are then factorized using NNSE.

For every word in our story, we therefore obtain 1000 NNSE features of which we keep the top 100 (these 100 features are picked from the 1000 based on the set of words in the story by choosing the dimensions with the highest average magnitude for these words, whereas the original 1000 were picked by the NNSE model based on the set of all words in the corpora). We sum the features of the four words within each TR.

Discourse features

Because of the difficulty of the problem of modeling the meaning of phrases computationally, we skipped for now the creation of IFSs that relate to sentence structure. We omitted IFSs related to non-literal meanings for the same reason. We also omit the description of "story grammar", i.e. annotating the beginning, conflict, climax of conflict and resolution of conflict, specifically because these each occur only once since only one story is read, and therefore it is not statistically feasible to dissociate their contribution to the signal from other confounding factors.

Because of how chapter 9 is structured, we decided to capture the narrative structure by using features that identify the different story characters and features that characterize the events they participate in: physical motions they perform, non-motion actions they perform and the emotions they experience. This chapter also contains frequent instances of directly quoted speech, and therefore we used the presence of dialog as a feature. Other narrative elements such as location did not vary or occur frequently in the chapter and therefore we excluded them. We made the following annotations manually by going through the story text:

- Characters: We resolve all pronouns to the character to whom they refer, and make binary features to signal which of the 10 characters are mentioned.

- Motions: We identified a set of motions that occurred frequently in the chapter (e.g. fly, manipulate, collide physically, etc.). Because the actions happen in the course of a sentence, we created two story features for: a punctual feature and a "sticky" feature. The punctual feature represented when the verb of the motion was mentioned, and the sticky feature is on for the duration of the motion (i.e. the sentence). We disregarded some of the punctual motion features because they had very few occurrences.

- Speech: We indicated the parts of the story that corresponded to direct speech between the characters. We have a punctual feature that indicates the verb that announces which character is speaking (e.g. "said Harry"), and a sticky feature that indicates ongoing speech.

- Emotions: We identified a set of emotions that were felt by the characters in the chapter (e.g. annoyance, nervousness, pride, etc.). We had punctual features for when the emotion was explicitly mentioned, and sticky features when it was being felt by the characters.

- Verbs: (non-motion) We identified a set of actions that occurred frequently in the chapter that were distinct from motion (e.g. hear, know, see, etc.). These typically spanned a shorter time than motions and we only used punctual features to represent them.

Semantics	1...100	Syntax	150 Sentence Length
Speech	101 speak - sticky	-parts of speech	151 ,
	102 speak - punctual		152 .
Motion	103 fly - sticky		153 :
	104 manipulate - sticky		154 Coordinating conjunction
	105 move - sticky		155 Cardinal number
	106 collide physically - sticky		156 Determiner
	107 fly - punctual		157 Preposition / sub. conjunction
	108 manipulate - punctual		158 Adjective
	109 move - punctual		159 Modal
Emotion	110 annoyed - punctual		160 Noun, singular or mass
	111 commanding - punctual		161 Noun, plural
	112 dislike - punctual		162 Proper noun, singular
	113 fear - punctual		163 Proper noun, plural
	114 like - punctual		164 Personal pronoun
	115 nervousness - punctual		165 Possessive pronoun
	116 questioning - punctual		166 Adverb
	117 wonder - punctual		167 Particle
	118 annoyed - sticky		168 to
	119 commanding - sticky		169 Interjection
	120 cynical - sticky		170 Verb, base form
	121 dislike - sticky		171 Verb, past tense
	122 fear - sticky		172 Verb, gerund or present part.
	123 mental hurting - sticky		173 Verb, past part.
	124 physical hurting - sticky		174 Verb, non-3rd person sing. present
	125 like - sticky		175 Verb, 3rd person sing. present
	126 nervousness - sticky		176 Wh-determiner
	127 pleading - sticky		177 Wh-pronoun
	128 praising - sticky		178 Wh-adverb
	129 pride - sticky	-dependency roles	179 Unclassified adverbial
	130 questioning - sticky		180 Modifier or adjective or adverb
	131 relief - sticky		181 Coordination
	132 wonder - sticky		182 Coordination
Verbs	133 be		183 Other dependent (default label)
	134 hear		184 Indirect object
	135 know		185 Modifier of noun
	136 see		186 Object
	137 tell		187 Punctuation
Characters	138 Draco		188 Modifier of preposition
	139 Filch		189 Predicative complement
	140 Harry		190 Parenthetical
	141 Hermione		191 Particle
	142 Mrs. Hooch		192 Root
	143 Mrs. McGonagall		193 Subject
	144 Neville		194 Verb chain
	145 Peeves		195 Modifier of verb
	146 Ron		
	147 Wood		
Visual	148 Average Word Length		
	149 Variance of Word Length		

Table 4.1: List of all the textual features.

	They were half hoping	for a reason to	fight Malfoy, but Professor	McGonagall, who could spot	...	Harry had heard Fred	and George Weasley complain	about the school brooms,	saying that some of
Semantic 1	0	0	0.12	0		0.13	0.11	0	0.01
speak - sticky	0	0	0	0		0	0	0	0
fly - sticky	0	0	0	0		0	0	0	0
manipulate - sticky	0	0	0	0		0	0	0	0
move - sticky	0	0	0	0		0	0	0	0
collide physically - sticky	0	0	0	0		0	0	0	0
hear	0	0	0	0		1	0	0	0
Draco	0	0	1	0		0	0	0	0
Filch	0	0	0	0		0	0	0	0
Harry	1	0	0	0		1	0	0	0
Hermione	0	0	0	0		0	0	0	0
Mrs. Hooch	0	0	0	0		0	0	0	0
Mrs. McGonagall	0	0	0	1		0	0	0	0
Average Word Length	4.5	3	6	5.75		4.25	6	5.25	4
Personal pronoun	1	0	0	0		0	0	0	0
Possessive pronoun	0	0	0	0		0	0	0	0
Object	0	0	1	0		0	1	0	0
Verb chain	1	0	0	1		1	0	0	0

Table 4.2: Example of the time course of the different types of story features for two story passages. Stories have to be represented in a feature space that allows for learning the brain response to individual features. The neural response to a novel part of the story can then be predicted as the combination of the responses associated with its features.

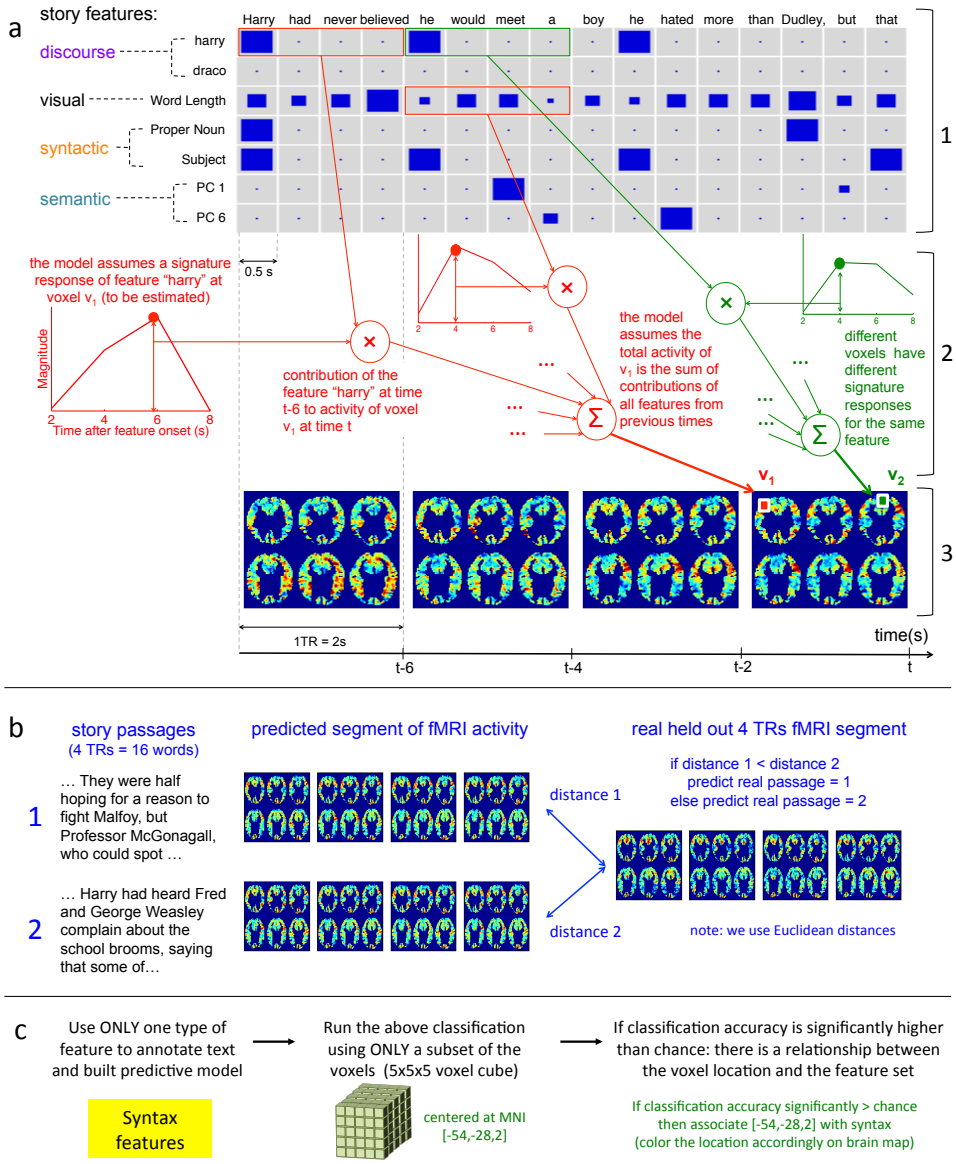
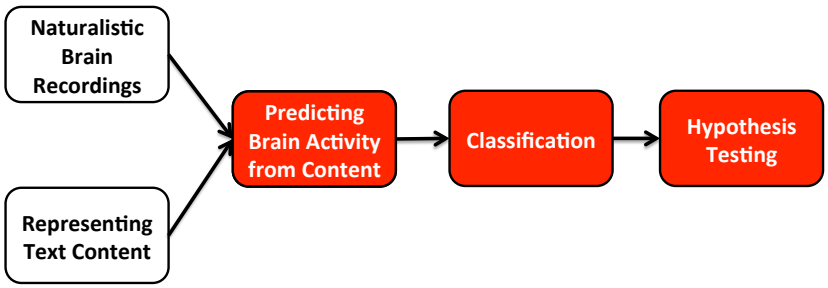


Figure 4.2: **Illustration of the model and the classification task. a**- (1) Diagram showing 7 of the 195 story features used to annotate a typical story passage. The size of each square indicates the magnitude of the feature. (2) Diagram of our generative model. The model assumes that the fMRI neural activity at each voxel at time t depends potentially on the values of every story feature for every word read during the preceding 8s. Parameters learned during training determine which features actually exert which influence on which voxels' activity at which times. A rectangle around 4 consecutive feature values indicates these values correspond to one time point and their magnitudes were summed. (3) Time course of fMRI volumes acquired from one subject while they read this specific story passage. Only 6 slices are shown per volume. **b**- Classification task. We test the predictive model by its ability to determine which of two candidate story passages is being read, given a time series of real fMRI activity held out during training. The trained model first predicts the fMRI time series segments for both of the candidate story passages. Then it selects the candidate story passage whose predicted time series is most similar (in Euclidean distance) to the held out real fMRI time series. **c**- Diagram of how we discover what type of information is processed by different regions. We repeat this procedure for every feature set and every location and we use the results to build representation maps.

4.3 Computational model and classification approach



We trained a computational model to predict the observed sequence of fMRI brain activity while the subjects read chapter 9 of *Harry Potter and the Sorcerer's Stone* [106]. This is a **continuous fMRI** modeling approach, as described in chapter 3. To characterize the input time series of text (of which each word was shown for 0.5s), a vector time series was created with 195 story features whose values change every 0.5s.

As described in the previous section, because one fMRI image is acquired every 2s, the model collapses the 0.5s time series of story feature vectors by summing the story feature vectors associated with the four consecutive words presented in each 2s interval. The result is a story features time series with values every 2s, aligned to the timing of the fMRI data acquisition.

The model predicts the neural activity at each voxel independently. It assumes that each time a particular story feature $x^{(j)}$ occurs, it will generate the same response signature in voxel v, weighted by that feature's value. Since changes in the fMRI signal persist for approximately 8s after neural activity and the signal is sampled with a period of 2s, the model estimates this response signature for feature x_j as a series of points $\{\beta_v^{1(j)}, \beta_v^{2(j)}, \beta_v^{3(j)}, \beta_v^{4(j)}\}$ corresponding respectively to times $\{2, 4, 6, 8\}$ seconds after feature onset. The signal in a voxel v at time t is therefore modeled as:

$$y_{vt} = \sum_{k=1}^{4} \mathbf{x}_{t-k}^{\top} \boldsymbol{\beta}_{v,k} + \epsilon_{vt}.$$

The activity at voxel v is the sum of the contributions of the F story features. Each feature x_j's contribution is the convolution of its magnitude over time with its temporal response signature at voxel v. This is illustrated in Fig. 3.2 and more details are listed in chapter 3.

Every voxel's activity at time t is thus a linear combination of all story features at the four preceding time points, where the specific linear combination is determined by the set of learned $\beta_v^{k(j)}$ parameters. ℓ_2-regularized linear regression was used to learn the very large set of parameters[2]. The model is trained independently for each subject in the study. Note the parameters $\beta_v^{1:4(j)}$ that represent a single time signature response are learned with no assumption on the shape of the response function, observed in fMRI time series. On average, we obtain for some types of features concave time series shapes that resemble the characteristic shape of the typical

[2]we pick the ℓ_2-penalty parameter independently for every voxel. We use generalized cross validation [36] to estimate the average leave one out cross validation error for each possible value of the penalty parameter.

fMRI hemodynamic response (Appendix B). However, our model also allows for the possibility that certain story features evoke very complex time series of neural activity whose fMRI signatures vary greatly from the standard hemodynamic response to a single isolated impulse of neural activity. Consequently, for some types of features, we learn more complex impulse responses. We have tried using more time points to estimate the response (5 and 6 instead of 4), however we did not find any region in which the model improved significantly in performance. Because we already have a large number of covariates (195 features $\times$ number of time windows) and a fixed number of samples, we chose to use 4 time points. Fig. 4.2(a) shows a summary of the predictive model.

Whole Brain Classification

To evaluate the model's accuracy, a cross-validation approach was used in which the model was repeatedly trained and tested. In each cross-validation fold, only 90% of the story time series and associated fMRI data were used for training the model, while the remaining 10% were held-out as test data. We divided the held-out story times series and the associated fMRI data into non-overlapping time-series segments of length 20 TRs. Fig. 4.2(b) summarizes how the accuracy of model predictions was assessed (in that figure, the segments are of length 4TRs for simplicity but the concept is the same). We go through the held out 20 TRs fMRI time series; for each one of the time-series, we perform a classification task that aims to identify the correct 20 TR story passage out of two possible choices (the corresponding 20 TRs passage and another one chosen at random). The classification is done in two steps. (1) The model predicts the fMRI time series for each of these two passages, for each of the human subjects in the study (recall that a different model is trained for each human subject). The predicted fMRI time series for all 8 subjects are then concatenated to form a predicted *group fMRI time series* covering all subjects in the study. (2) The held out group fMRI time series (which also corresponds to the concatenation of the 8 subjects' time-series) is then compared to the two predicted group time series and the model is required to determine which of the two passages was being read when the observed group fMRI data was collected. To answer this two-choice classification task, the model chooses the passage whose predicted group fMRI time series is closest (in Euclidean distance) to the observed group fMRI time series.

Note that the chance-level performance in this two-way classification of text passages over the held-out data is 50%. Also note that both the learning and classification steps were done without averaging data over subjects or making assumptions on their brain alignment. Further details are provided in chapter 3. Finally, note that we repeat the classification of each fMRI segment a large number of times with different alternative choices to minimize the variance of the results. The boundaries of the passages we choose are arbitrary since the selection is made automatically and all of the story passages are constrained to be of the same size, i.e. the two test passages do not correspond to defined paragraphs or sections of the text. Because we pair each true passage with many other passages in different classification tasks and average the accuracy over all the tasks, we minimize confounds that might occur because two specific passages are extremely different in some way that is tangent to the information content we are studying.

Uncovering Different Patterns of Representation

We wished to explore which story features mapped to which locations in the brain. To find this mapping the above classification approach was followed, but using **only one type of story feature at a time** to annotate the text passage (e.g. only the semantic features). Fig. 4.2(c) describes this approach. We also limited the predictions to a **small subset of the voxels** in a Searchlight-like [65] manner that we call concatenated Searchlight. This concatenated Searchlight uses a $15 \times 15 \times 15$mm cube centered at one voxel location (corresponding to $5 \times 5 \times 5$ voxels). After normalizing the subjects to the MNI (Montreal Neurological Institute) space, we include in each cube the set of voxels from all subjects whose coordinates fall into the cube (subjects may differ in how many voxels they contribute to a particular cube because of the disparity in the size of their ventricles or the shape of the surface of their brain).

Our concatenated Searchlight is not equivalent to spatial or cross-participant smoothing because, again, the voxels associated with each subject are treated independently. The difference is discussed in Appendix D. Because the voxel cube used is larger than one voxel ($5 \times 5 \times 5$ voxels), this method searches for regions with high accuracy across subjects while allowing for small anatomical variations among their brains.

By successively testing every **type of feature** j at every cube **location** r, we determine in which brain regions each type of feature yields high classification accuracy. Our assumption is that, if using feature set j in location r yields a high classification accuracy, then the activity in region r is modulated by feature set j, i.e. region r represents feature j. For example, if using part of speech features allows us to classify very accurately a region in the temporal pole, then this suggests that this region of the temporal pole is representing part of speech information.

To assess the significance of the classification accuracies an empirical distribution of chance level performance was estimated. We then corrected for multiple comparisons (using the method described in chapter 3). From the classification results, we therefore obtain accuracy maps that allow us to determine where each type of information is represented by fMRI activity.

4.4 Results and Discussion

Whole Brain Classification Results

We compute the average classification accuracy of our model when predicting fMRI time series associated with text passages that were not observed during training. The model is able to classify which of two novel passages of the story is being read with an accuracy of 74%. This is significantly higher than chance accuracy, which is 50% in this balanced task ($p < 10^{-8}$), indicating that the model can indeed distinguish between the literary content of two novel text passages based on neural activity while these passages are being read.

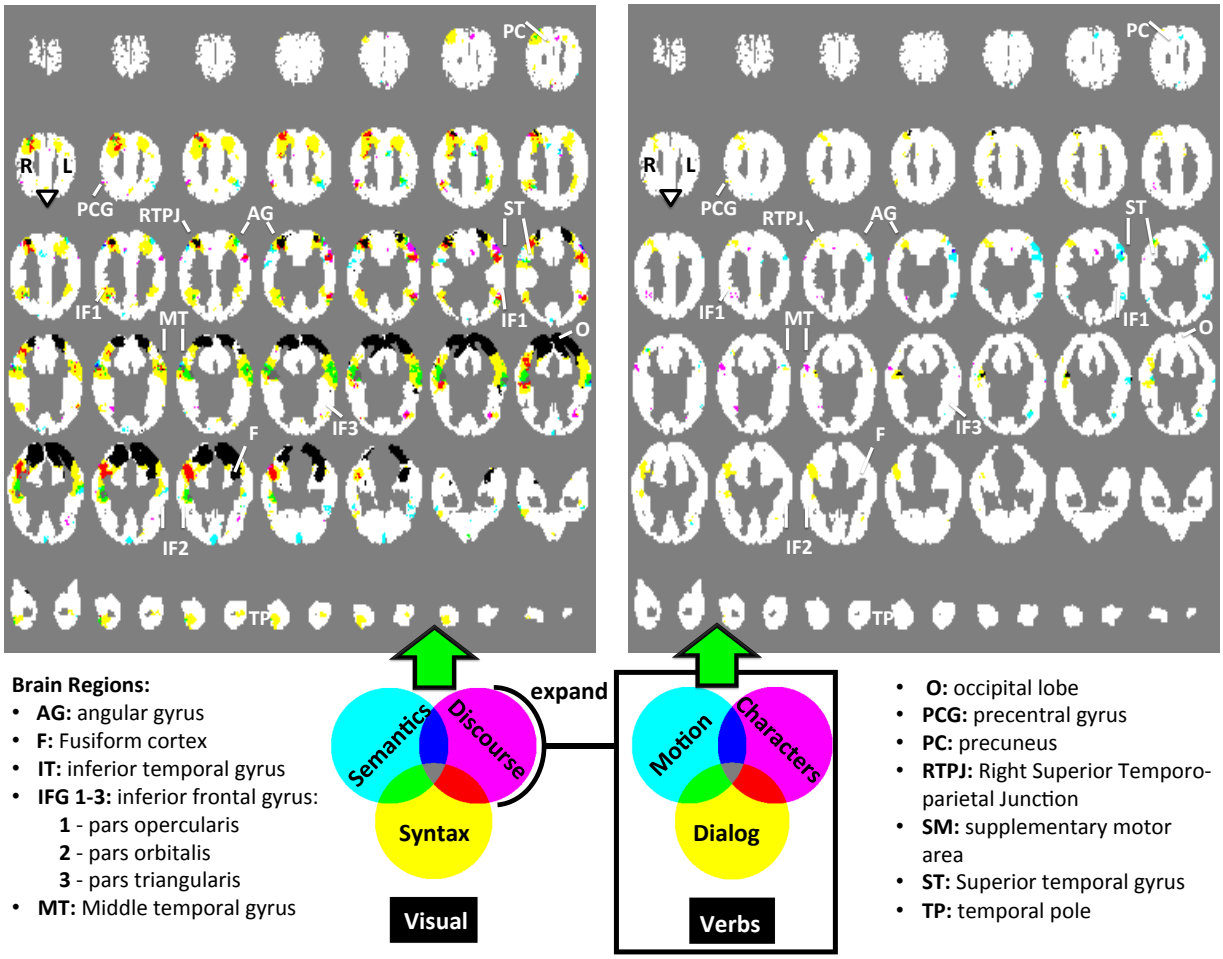


Figure 4.3: **Accuracy maps revealing different patterns of representation of different reading processes. (Left)** Voxels with significantly higher than chance classification accuracy when using different types of story elements as features, shown in different colors corresponding to the type of story elements. The brain used here is a superset of the brain of the 8 subjects, i.e. the union of all the voxel locations in the 8 brains. The slices are drawn such that they increase in the Z MNI-coordinate when going right to left, then bottom-up. Within each slice, the top of the slice corresponds to the posterior of the brain, and the right side of the slice corresponds to the left side of the brain. Each voxel location represents the classification done using a cube of $5 \times 5 \times 5$ voxel coordinates, centered at that location, such that the union of voxels from all subjects whose coordinates are in that cube are used. **(Right)** Voxels with significantly higher than chance classification accuracy when using different types of discourse elements as features, shown in different colors corresponding to the type of discourse elements.

The successful classification results we obtain indicate that, despite the low temporal resolution, it is possible to investigate the fast dynamic process of reading at a close-to-normal pace using fMRI, and to train a computational model of story comprehension that can successfully predict the time series of neural fMRI activity generated by reading novel passages of text. This model tracks multiple levels of processing of the story and links them to different brain areas. Our approach combines data from multiple subjects while allowing for subject-to-subject anatomical variability, makes minimal assumptions about the shape of the time series response to different story features in different brain regions, and learns the shape of these responses from observed data. As an extra advantage, authentic stories provide engaging experimental stimuli which helps subjects to remain alert.

We set out next to investigate how the different types of cognitive processes that underlie story reading are represented in the brain. For that purpose we ran the concatenated Searchlight approach, described in the methods section, using different input features and we constructed representation maps, which we discuss next.

Different Patterns of Representation

Fig. 4.3(Left) shows the map of statistically significant classification accuracy (controlled at a false discovery rate of $\alpha = 0.05$) for the four categories of story features: semantics, syntactic, discourse features and visual features. Fig. 4.3(Right) offers a closer look at the different categories of discourse features. Fig. 4.4(b) shows the learned map on the surface of a brain template. We did not find regions with significantly higher than chance accuracy along the medial wall and therefore we don't show it in Fig. 4.4(b). We discuss the different regions in this section.

Word Length We find that the regions from which we can decode using the word length properties are in the occipital cortex, spanning the visual cortex (V1-4,VO1-2). This result is highly expected, and serves as an initial sanity check since the regions with high classification accuracy are mainly in the visual cortex. The visual regions are larger in the left hemisphere, spreading to the left fusiform cortex. This is most probably due to the activity of the Visual Word Form Area [15] that is being modulated by word length.

Syntax and Structure Our results indicate that multiple areas in the brain represent language structure and syntax. Some of these regions are expected while others are somewhat surprising. Our syntax and structure features were composed of features related to part of speech and punctuation, grammatical role of a word in a sentence and the ordinal number of the word in the sentence. These features therefore capture a rich array of information: they are not only a measure of syntactic complexity but they also capture the different grammatical structures of the sentences in the text.

Fig. C.1 in Appendix C shows the breakdown of the syntax regions along our three types of features. In [95] the authors identified a network of regions where neural activity was correlated with the length of linguistic constituents. Using the sentence length feature, we were able to recover only the left temporo-parietal region that is reported (when using non-smoothed data - see Appendix C - we are also able to recover the left posterior superior temporal sulcus region

that is reported). Interestingly, we find many more regions in the right temporo-parietal cortex that are related to sentence length. These regions are also modulated by the other syntactic features as well as by the presence of dialog. This indicates that these regions are modulated by the complexity and length of sentences. The right parietotemporal cortex has been implicated previously in verbal working memory processes [104] and has been shown to be more activated for good readers than for poor readers [75].

The strong right temporal representation of syntax that we found was not expected. Indeed we did not find papers that report the large right hemisphere representation of sentence structure or syntax that we obtain. One reason might be that our syntax features are unique: whereas most experiments have approximated syntactic information in terms of processing load (length of constituents, hard vs easy phrase structure etc.) we model syntax and structure using a much more detailed set of features. Specifically, our model learns distinct neural encodings for each of 46 detailed syntax features including individual parts of speech, (adjectives, determiners, nouns, etc.) specific substructures in dependency parses (noun modifiers, verb subjects, etc.), and punctuation. Earlier studies considering only increases or decreases in activity due to single contrasts in syntactic properties could not detect detailed neural encodings of this type. We hypothesize that these regions have been previously overlooked.

The regions we find in the bilateral temporal cortices are related to both dependency role and part of speech features, indicating that they might be involved in both integration of the successive words and the representation of the incoming words. regions that are slightly more posterior represent part of speech features (features of the incoming words) and the ones that are slightly more anterior represent dependency roles (i.e. are implicated in word integration and sentence structure building). Regions in the bilateral temporal poles and the right IFG are representing dependency roles, indicating more high level processing, while the left IFG represents both dependency roles and parts of speech.

Lexical Semantics Our model also found parts of the brain that represent semantics of individual words. Some of these areas such as the left superior and middle temporal gyrii and the left IFG have frequently been reported by others to represent semantics during language processing [45]. We found a right middle temporal representation of semantics. This is consistent with a theory of coarse semantic representation in the right hemisphere [74]. We also found semantic representation in the medial frontal cortex as well as the bilateral angular gyrii and the left pre-central gyrus.

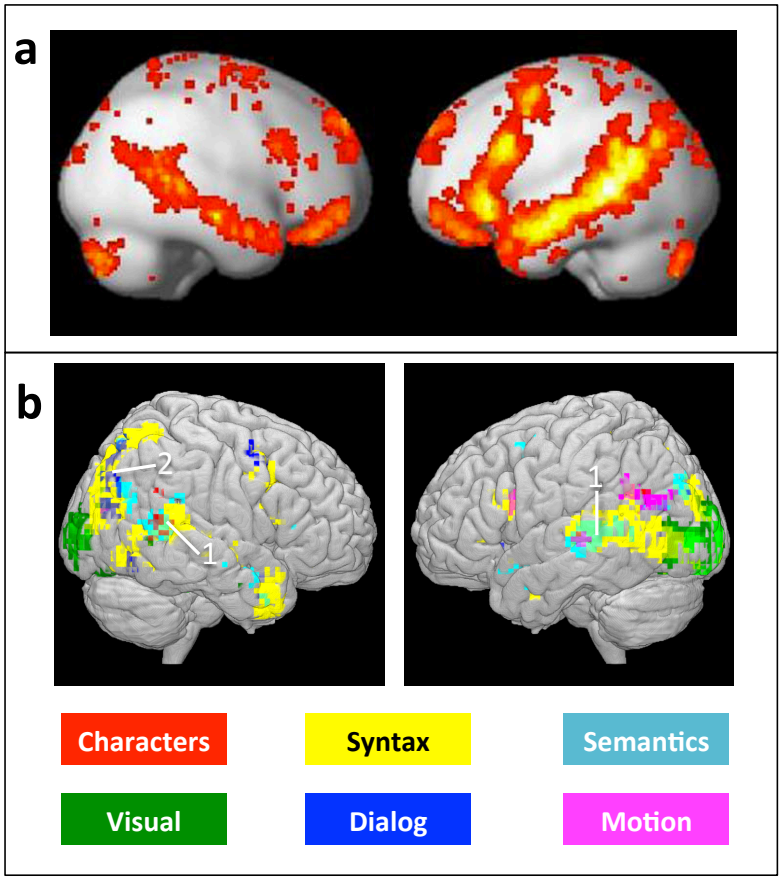


Figure 4.4: **Map of the patterns of representation compared with the regions involved in sentence processing: our method recovers similar regions and differentiates them according to which information process they represent.** **a**- Adapted from [23] (Fig. 1): "The language system: a set of brain regions that are robustly and consistently activated by linguistic input (see [24], for further discussion of how to define the "language system/network"). A probabilistic activation overlap map for the contrast between sentences and sequences of pseudowords (adapted from [22]). Warmer colors indicate greater proportions of subjects showing a reliable sentences $>$ pseudoword lists effect." **b**- Results obtained by our generative model, showing where semantic, discourse, and syntax information is encoded by neural activity. Note this model identifies not just where language processing generates neural activity, but also what types of information are encoded by that activity. Each voxel location represents the classification done using a cube of $5 \times 5 \times 5$ voxel coordinates, centered at that location, such that the union of voxels from all subjects whose coordinates are in that cube are used. Voxel locations are colored according to the feature set that can be used to yield significantly higher than chance accuracy. Light green regions, marked with (1), are regions in which using either semantic or syntactic features leads to high accuracy. Dark gray regions, marked with (2), are regions in which using either dialog or syntactic features leads to high accuracy.

Dissociation of Syntax and Semantics The question whether the semantics and syntactic properties are represented in different location has been partially answered by our results. There seems to be a large overlap in the areas in which both syntax and semantics are represented. This is partially in alignment with what [22] found. The authors found that all the regions responsive to language stimulus were responsive to both syntax and semantics information. They were however able to distinguish between pure semantic information (word lists) and pure syntactic information (Jaberwocky) in some of the regions, leading them to conclude that in some of the regions syntactic and semantic information were not very closely represented and could be distinguished by voxel activity. They also found the lexical semantic information to be more strongly represented than the syntactic information. Using our natural story reading paradigm, we have found partially similar results: many regions in the bilateral temporal cortices seem to be coding both semantic and syntactic meaning, leading to one of two conclusions: either these brain regions process a meaning that is common to semantic and syntactic properties of words that are closely linked together, or our features are themselves representing information at the intersection of semantics and syntax that is related to the activity in that region. Furthermore, we find (1) regions that are selectively processing syntax and semantics and (2) that syntactic information is more widely and strongly represented. The difference could be due to the richness of our syntactic features and the additional fact that they indirectly measure verbal working memory and effort, which would recruit general purpose areas that exceed the language network.

Discourse and narrative features Our results reveal a variety of brain regions that encode different information about story characters. Physical motions of story characters were represented in the posterior temporal cortex/angular gyrus, a region implicated in the perception of biological motion [44]. It has been shown that imagined biological motion also activates this area [44]. Processing the motions of the characters also modulated the activity of a region in the superior temporal sulcus, as well as in the left inferior frontal gyrus.

Presence of dialog among story characters was found to modulate activity in many regions in the bilateral temporal and inferior frontal cortices; one plausible hypothesis is that dialog requires additional processing in the language regions. More interestingly, it seems like presence of dialog activates the right temporo-parietal junction, a key theory of mind region [109]. This observation raises an exciting hypothesis to pursue: that the presence of dialog increases the demands for perspective interpretation and recruits theory of mind regions.

The identities of different story characters can be distinguished based on neural activity in the right posterior superior/middle temporal region. In [74] a "protagonist's perspective interpreter network" is outlined, based on a review of multiple studies. It encompasses among others the right posterior superior temporal gyrus. This region is also a classical theory of mind area [109], and has been found to encode facial identity [34].

Differentiation of areas and stability of results

We therefore find a different representation for each type of features, with somewhat little specificity of the individual language regions. We suspect that these results, while revealing if considered at a coarse spatial scale, are however dependent on the analysis approach when the exact voxel locations are desired. To illustrate this point, we show in Fig. C.2 and C.3 in Appendix C

the results from running the same model as ours, with the change that the data was not smoothed spatially beforehand. There is a large variation in the boundaries of the regions, while the main general locations have some consistency.

The reason for the difference in the results is that our classification method relies on ridge regression and learns a different penalty parameter for each voxel. This leads to learning very high penalty parameters for noisy voxels, and very small ones for good voxels, effectively resulting in an automatic voxel selection (chapter 3). When the data is spatially smoothed, this disturbs the voxel selection, reduces the selection effect and brings down accuracy slightly, resulting in a smoother thresholding and more interpretable map such as the one in Fig. 4.4 and C.1. It is however not straightforward to decide which method leads to more accurate spatial localization results. This observation really reveals the fickleness of brain imaging results, which are a general problem in the field, and their high dependence on even the analysis methods, which lead to different conclusions, especially when dealing with questions like specificity of regions. Analysis methods vary considerably between experiments, and it's not always clear which approach is more appropriate since multiple approaches can be statistically sound. This points to the urgency for establishing better standards and better methods that would be robust to such changes. We are currently working towards this goal.

An additional concern when looking at the regions identified for different features is that significance thresholding doesn't take into account that these different types of features have different statistical properties that influence their performance, and comparing them on the same metric introduces some arbitrariness. We discuss these issues in Appendix C, tables C.1 and C.2, and we show in Fig. C.4 a map in which we color the top 1000 voxels per feature in terms of accuracy, instead of coloring the voxels that exceed the significance threshold.

A comprehensive study of language processing

We have used our model to shed new light on what information is encoded by neural activity in different regions of the brain during story comprehension. Whereas previous research has shown which brain regions exhibit *increased brain activity* associated with different aspects of language processing, our results reveal in addition which brain regions *encode specific information* such as the identity of specific story characters. In recent research [22], a network of regions involved in language processing is obtained. It includes regions from the left angular gyrus to the left temporal pole, multiple left IFG regions, and multiple right temporal regions. That network is show in Fig. 4.4(a). Our own analysis, shown in Fig. 4.4(b), largely agrees with these findings, in terms of which regions exhibit language-related activity. However, as shown in Fig. 4.4(b), our analysis also reveals which of that neural activity is modulated by (and may therefore encode) specific perceptual, syntactic, semantic and discourse story features. Whereas previous work has studied some of these correspondences in isolation, the results presented here are the first to examine neural encodings of diverse story information at such a scale and across the brain in a realistic, story reading setting.

As illustrated in the above discussion, the model of reading introduced here can be used to study many aspects of reading simultaneously, without needing to vary just one dimension of the experimental stimulus at a time. This departure from the classical experimental setting has many advantages. We can use natural texts as stimuli, and study close-to-normal reading with

its natural diversity of language constructs and attendant neural subprocesses. This model is also very flexible – given a rich enough stimulus, one can add additional stimulus features that one wishes to study. As suggested by [84], one could analyze an experiment with a new set of features without needing to collect new brain image data for each phenomenon of interest.

The rise of brain image decoding has already made the neuroimaging field aware of the difference between (a) approaches that use the presence/absence of a stimulus and (b) approaches that use the presence of different instantiations of the stimulus. For example[88] distinguishes between regions that identify the presence of faces and regions that process the characteristics of faces. Out of the regions that are modulated by the presence of a face, the authors determine which regions can be used by a classifier to decode face identity. Using different instantiations of a stimulus (e.g. of a face) therefore allows us to find regions that encode the properties of the stimulus in consideration. In our experiment, we take this approach to the next level: there is only one stimulus (text) that is always being presented, and it is instantiated with a very large diversity (variations along a large number of dimension). More work is needed to understand more deeply how the different approaches of studying language tie together; and to understand how to combine what we can learn from experiments that rely on modeling the features of the stimuli (such as ours) versus experiments that contrast different types of information load (for example comparing stories to scrambled sentences and scrambled words such as in [59]). A compelling question that we have yet to answer is how much can we rely on modeling experiments, and how much can we stray from using controlled experiments. The similarity between our results and the literature we cited, and the fact that we reproduced many of these results using one modeling experiment only, are an encouraging first answer.

Furthermore, under the uncontrolled setting of our experiment, more work is needed in order to discount the effect of the correlation between the features sets. We obtain many regions that are related to multiple types of features, and it is crucial for our modeling approach to determine which of these associations are only due to the correlations between the feature sets. We are currently working on this problem and on expanding the computational methods we described here to give a clearer picture of the relationship between types of features and brain regions.

While the above discussion focuses on a map of group-wide language processing obtained from multiple subjects, it is also possible to use this approach to produce subject-specific reading maps. We suggest that our approach may be useful in the future to investigate language processing in a way that was not possible before. For example, one might test a hypothesis about how aphasic patients develop alternative processing routes by discovering the information encoded in each region for participants with aphasia and comparing the resulting distributions to controls. Similarly, subject-specific reading maps might be used to understand the cause of an individual's reading difficulties, and to better understand individual differences in reading processes. A further potential use is for pre-surgical mapping: this approach might help to identify, in parallel and with great precision, the patient-specific network of regions involved in language processing.

Chapter 5

The Timeline of Meaning Construction

Language processing is a fast dynamic process. Adults read words at a speed of 3 words a second on average, and at this fast rate, individual words are perceived and they are combined together to understand the meanings and structure of sentences. This complex process happens much faster than the slow acquisition time of fMRI[1]. In this chapter, we therefore turn to Magnetoencephalography (MEG), which offers a direct measure of neural activity that has very high time sensitivity. We will use this method to investigate how the brain perceives and understands the properties of consecutive words while reading and how it combines the words together and maintains a representation of this evolving context.

We follow here the Intermediate Feature Space (IFS) approach introduced in chapter 3. To study how the context of previous words and the properties of next word are represented, we need a numerical vector representation of the context created by consecutive words and the properties of individual words. In order to obtain such a representation, we noticed a parallelism between the brain and Recurrent Neural Network Language Models (RNNLM, proposed by [77]). An RNNLM is used to predict the next word given the previous series of words in a passage. To perform this task, this neural network has 3 key constituents: a **context vector** that summarizes the history of the previous words, a **properties vector** that summarizes the (constant) properties of a given word and finally the output **probability of a word** given the context. We use these vectors as IFSs in order to detect in the MEG data the brain processes that we are interested in, i.e. respectively: the brain representation for the **context relating to the previous words**, the **progressive perception of the incoming word** and the **integration process of this word with the previous words**.

This chapter is based on work we published in [128], in which we compare the RNNLM - which uses the entire history of words to model context - with Neural Probabilistic Language Models (NPLM) - which use limited context constrained to the recent words (3 grams or 5 grams) - and we find that both these models perform similarly at predicting brain activity. The vectors we obtain from these models allow us to find brain signatures for the corresponding processes of interest. In order to perform these experiments, we trained these models on a large Harry Potter fan fiction corpus and we then used them to predict the words of chapter 9 of *Harry Potter and the Sorcerer's Stone* [106]. In parallel, we ran an MEG experiment, which we extend here from the

[1]Remember fMRI has a sampling rate of a few seconds and is recording a slowly varying, indirect consequence of neural activity.

word in [128] to include 9 participants. These participants read the words of chapter 9 one by one while their brain activity was recorded using MEG. We then looked for the alignment between the word-by-word vectors produced by the neural networks and the word-by-word neural activity recorded by MEG.

We showcase here the alignment we were able to obtain between the RNNLM vectors and MEG activity. It allowed us to uncover a potential spatio-temporal course for context and new word representation in the brain. We also extend our IFS analysis from chapter 3 into a method that aims to test for the conditional independence of brain activity in region r and IFS j, given the other IFSs. We present these details in section 5.5. Different IFSs can be correlated, therefore if we find using our traditional method a relationship between brain area and r an IFA j, it might be exclusively due to the correlation between j and another IFS k, which happens to be the correct IFS that is related to brain area r. Our conditional independence test tries to find areas that has a relationship with IFS j that is not explained by the other IFSs in our set.

Using our test, we single out the contribution to brain activity of the current word, from the contribution of the previous words. Our results suggest each new word generates additional neural activity that encodes information about this word, flowing generally from visual posterior to left temporal, and to increasingly anterior/frontal brain regions. We use different IFSs we presented in chapter 4 in order to offer some first answers on what type of information (visual, syntactic or semantic) is being processed in various location during this progression.

5.1 Neural processes involved in reading

Reading requires us to perceive incoming words and gradually integrate them into a representation of the meaning. As words are read, it takes 100ms for the visual input to reach the visual cortex. 50ms later, the visual input is processed as letter strings in a specialized region of the left visual cortex [108]. Between 200-600ms, the word's semantic properties are processed. Less is understood about the cortical dynamics of word integration, as multiple theories exist [26, 46].

As we mentioned in chapter 2, magnetoencephalography (MEG) is a brain-imaging tool that is well suited for studying language. MEG records the change in the magnetic field on the surface of the head that is caused by a large set of aligned neurons that are changing their firing patterns in synchrony in response to a stimulus. Because of the nature of the signal, MEG recordings are directly related to neural activity and have no latency. They are sampled at a high frequency (typically 1kHz) that is ideal for tracking the fast dynamics of language processing.

In this chapter, we are interested in the mechanism of human text understanding as the meaning of incoming words is fetched from memory and integrated with the context. As mentioned previously, this is analogous to neural network models of language that are used to predict the incoming word. The mental representation of the previous context is analogous to the latent layer of the neural network which summarizes the relevant context before seeing the word. The representation of the meaning of a word is analogous to the property vectors that the neural network learns in training and then uses. Finally, one common hypotheses is that the brain integrates the word with inversely proportional effort to the surprisal of a word[2] [25]. There is a well studied

[2]i.e. how surprising the word is from the previous context, as measured by its conditional probability of occurring given the previous words.

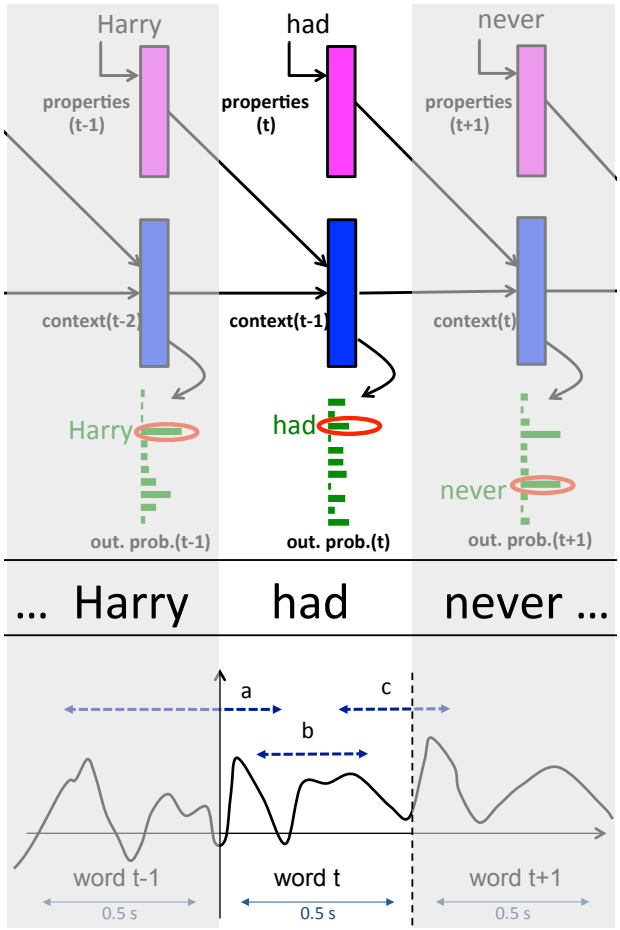


Figure 5.1: [Top] Sketch of the updates of a neural network reading chapter 9 after it has been trained. Every word corresponds to a fixed properties vector (magenta). A context vector (blue) is computed before the word is seen given the previous words. Given the context vector, the probability of every word can be computed (symbolized by the histogram in green). We only use the output probability of the actual word (red circle). [Bottom] Hypothetical activity in an MEG sensor when the subject reads the corresponding words. The time periods approximated as a, b and c can be tested for information content relating to: the context of the story before seeing word t, the representation of the properties of word t and the integration of word t into the context (the output probability of word t). The periods drawn here are only a conjecture on the timings of such cognitive events.

response known as the N400: it is an increase of the activity in the temporal cortex due to semantically incongruous stimulus. It has lately been shown that the N400 is not an all or none process that appears only in incongruous case, but that it is graded by the amount of surprisal of the incoming word given the context[3]. This is analogous to the output probability of the incoming word from the neural network.

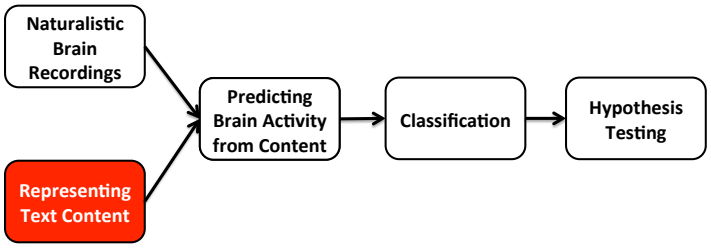
Fig. 5.1 shows a hypothetical activity in an MEG sensor as a subject reads a story in our experiment, in which words are presented one at a time for 500ms each. We conjecture that the activity in time window a, i.e. before word t is understood, is mostly related to the previous context before seeing word t. We also conjecture that the activity in time window b is related to understanding word t and integrating it into the context, leading to a new representation of context in window c.

Using three types of features from a recurrent neural network language model (context rep-

[3]In [25], the amount of surprisal that a word has given its context is used to predict the intensity of the N400 response. This is the closest study we could find to our approach. This study was concerned with analyzing the brain processes related only to surprisal while we propose a more integral account of the processes in the brain. The study also didn't address the contribution we propose here, which is to shed light on the inner constituents of language models using brain imaging. This study, as is mostly done when analyzing the N400 response, only focus on the N400 response for target words typically at the end of sentences. However, we model the surprisal effect for all words in a naturalistic text.

resentation, output probabilities and word properties), we therefore set to predict the activity in the brain in different time windows. We want to align the brain data with the various model constituents to understand where and when different types of processes are computed in the brain. An interesting extension in the future would be to use the brain data to improve the learning of neural network models and shed light on what their uninterpretable vectors are representing.

5.2 Recurrent Neural Network Language Model



Similar to standard language models, neural language models also learn probability distributions over words given their previous context. However, unlike standard language models, words are represented as real-valued vectors in a high dimensional space. We refer to these word vectors as *properties vector* (they are also sometimes referred to as word embeddings) and are learned from training data. Thus, although at training and test time, the input and output to the neural language models are *one-hot*[4] representation of words, it is their properties vector that are used to compute word probability distributions. After training the properties vectors are fixed and it is these vectors that we will use later on to predict MEG data. To predict MEG data, we will also use the latent vector representations of context that these neural networks produce, as well as the probability of the current word given the context. In this section, we will describe how RNNLMs compute word probabilities. We have also used feedforward NPLM models from [123] to predict MEG activity. However, because the results were very similar, we omit them from this document. These results are described extensively in [128].

Because we have applied the RNNLM exactly as is was introduced by [77], and using the toolkit provided by the authors, we only include here the high level description of this model. For more details, the reader is requested to refer to the original paper describing the model.

Unlike standard feedforward neural language models that only look at a fixed number of past words, recurrent neural network language models use all the previous history from position 1 to $t-1$ to predict the word at time t. This is typically achieved by *feedback* connections, where the hidden context layer activations used for predicting the word in position $t-1$ are fed back into the network to compute the hidden context layer activations for predicting the next word. The hidden layer thus stores the history of all previous words. We use the RNNLM architecture as described in [76], shown in Figure 5.2. The input to the RNNLM at position t is the one-hot representation of the current word, $\mathbf{w}(t)$, and the activations from the hidden layer at time $t-1$, $\mathbf{s}(t-1)$. The value of the hidden layer units at time $t-1$ is

[4]Indicator vector of the size of the vocabulary with all entries equal to 0, except the one corresponding to the word, which is equal to 1.

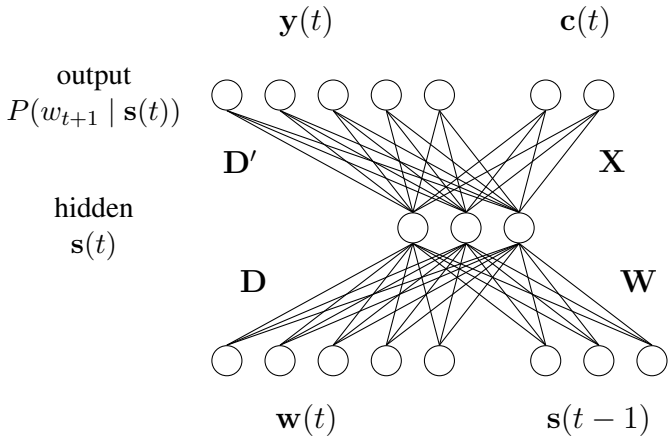


Figure 5.2: Recurrent neural network language model.

$$\mathbf{s}(t) = \phi\left(\mathbf{D}\mathbf{w}(t) + \mathbf{W}\mathbf{s}(t-1)\right),$$

where $\mathbf{D}$ is the matrix of input word properties, $\mathbf{W}$ is a matrix that transforms the activations from the hidden layer in position $t-1$, and ϕ is a sigmoid function, defined as $\phi(x) = \frac{1}{1+\exp(-x)}$, that is applied elementwise. The network is trained to compute the probability of the next word $\mathbf{w}(t+1)$ given the hidden state $\mathbf{s}(t)$. For fast estimation of output word probabilities, [76] divides the computation into two stages: First, the probability distribution over *word classes* is computed, after which the probability distribution over the subset of words belonging to the class is computed. The number of classes is fixed in advance and the words are automatically assigned to classes based on their unigram frequencies, to make sure frequent words are evenly divided among classes. The probability of a particular class with index m at position t is computed as:

$$P(\mathbf{c}_m(t) \mid \mathbf{s}(t)) = \frac{\exp\left(\mathbf{s}(t)\mathbf{X}\mathbf{v}_m\right)}{\sum_{c=1}^{C}\left(\exp\left(\mathbf{s}(t)\mathbf{X}\mathbf{v}_c\right)\right)},$$

where $\mathbf{X}$ is a matrix of class properties and $\mathbf{v_m}$ is a one-hot vector representing the class with index m. The normalization constant is computed over all classes C. Each class specifies a subset V' of words, potentially smaller than the entire vocabulary V. The probability of an output word l at position $t+1$ given that its class is m is defined as:

$$P(\mathbf{y}_l(t+1) \mid \mathbf{c}_m(t), \mathbf{s}(t)) = \frac{\exp\left(\mathbf{s}(t)\mathbf{D}'\mathbf{v}_l\right)}{\sum_{k=1}^{V'}\left(\exp\left(\mathbf{s}(t)\mathbf{D}'\mathbf{v}_k\right)\right)},$$

where $\mathbf{D}'$ is a matrix of output word properties and $\mathbf{v}_l$ is a one hot vector representing the word with index l. The probability of the word $\mathbf{w}(t+1)$ given its class c_i can now be computed as:

$$P(\mathbf{w}(t+1) \mid \mathbf{s}(t)) = P(\mathbf{w}(t+1) \mid c_i, \mathbf{s}(t))P(c_i \mid \mathbf{s}(t)).$$

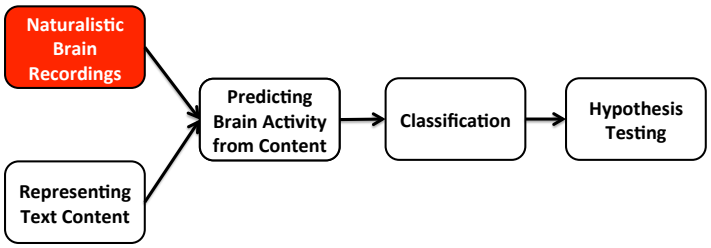
Training the Neural Network

We used the freely available training tools provided by [76][5] to train our RNNLM model used in our brain data classification experiments. Our training data comprised around 67.5 million words for training and 100 thousand words for validation from the Harry Potter fan fiction database (http://harrypotterfanfiction.com). We restricted the vocabulary to the top 100 thousand words which covered all but 4 words from Chapter 9 of *Harry Potter and the Sorcerer's Stone*.

We trained models with different hidden layers and learning rates and found the RNNLM with 250 hidden units to perform best on the validation set. We extracted our word properties from the input matrix $\mathbf{D}$ (Fig. 5.2). We used the default settings for all other hyperparameters.

5.3 Approach

We describe in this section our approach. In summary, we trained the neural network models on a Harry Potter fan fiction database. We then ran these models on chapter 9 of *Harry Potter and the Sorcerer's Stone* [106] and computed the context and propertied vectors and the output probability for each word. In parallel, 8 subjects read the same chapter in an MEG scanner. We build models that predict the MEG data for each word as a function of the different neural network constituents. We then test these models with a classification task that we explain below. We detect correspondences between the neural network components and the brain processes that underlie reading in the following fashion. If using a neural network vector (e.g. the RNNLM properties vector) allows us to classify significantly better than chance in a given region of the brain at a given time (e.g. the visual cortex at time 100-125ms), then we can hypothesize a relationship between that neural network constituent and the time/location of the analogous brain process.



5.3.1 MEG paradigm

We recorded MEG data for 9 subjects (5 female and 4 male) while they read chapter 9 of *Harry Potter and the Sorcerer's Stone* [106] (one subject's data had very strong noise and artifact and we did not use it, leaving 8 subjects). The participants were native English speakers and right handed. They were chosen to be familiar with the material: we made sure they had read the Harry Potter books or seen the movies series and were familiar with the characters and the story.

[5] `http://rnnlm.org/`

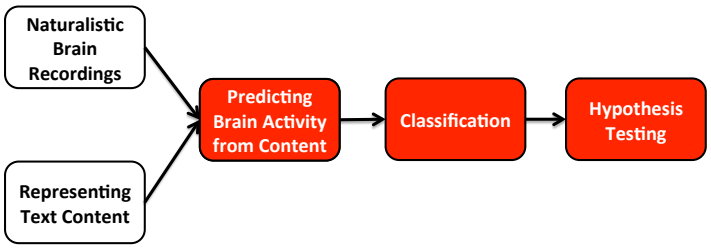
All the participants signed the consent form, which was approved by the University of Pittsburgh Institutional Review Board, and were compensated for their participation.

The words of the story were presented in rapid serial visual format [12]: words were presented one by one at the center of the screen for 0.5 seconds each. The text was shown in 4 experimental blocks of $\sim$11 minutes. In total, 5176 words were presented. Chapter 9 was presented in its entirety without modifications and each subject read the chapter only once[6].

One can think of an MEG machine as a large helmet, with sensors located on the helmet that record the magnetic activity. Our MEG recordings were acquired on an Elekta Neuromag device at the University of Pittsburgh Medical Center Presbyterian Hospital. This machine has 306 sensors distributed into 102 locations on the surface of the subject's head. Each location groups 3 sensors or two types: one magnometer that records the intensity of the magnetic field and two planar gradiometers that record the change in the magnetic field along two orthogonal planes[7].

Our sampling frequency was 1kHz. For preprocessing, we used Signal Space Separation method (SSS, [120]), followed by its temporal extension (tSSS, [119]).

For each subject, the experiment data consists therefore of a 306 dimensional time series of length $\sim$45 minutes. We averaged the signal in every sensor into 25ms non-overlapping time bins. Since words were presented for 500ms each, we therefore obtain for every word $p = 306 \times 20$ values corresponding to 306 vectors of 20 points.



5.3.2 Decoding experiment

To find which parts of brain activity are related to the three sets of word features derived from the neural network (the RNNLM context vector, the word properties vector or the probability of each word), we use the IFS approach described in chapter 3. We run a prediction and classification experiment in a 10-fold cross-validated fashion. At every fold, we train a linear model to predict MEG data as a function of one of the three IFS, using 90% of the data. On the remaining 10% of the data, we run a classification experiment.

Annotation of the stimulus text We have 3 sets of IFS annotations for the words of the experiment. Each set j can be described as a matrix $\mathbf{F}_j$ in which each row t corresponds to the vector

[6]For four of these subjects, we also acquired data for chapter 10, however, we have not yet used this data or analyzed it.

[7]In this chapter, we treat these three different sensors as three different dimensions without further exploiting their physical properties.

of annotations of word t of type j. As mentioned above, our three types of word annotations are derived from the RNNLM:

- the representation of the context before word t, i.e. $\mathbf{s}(t-1)$
- the output probability of word t, i.e. $P(\mathbf{w}(t)|\mathbf{s}(t-1))$
- the fixed properties for word t, i.e. $\mathbf{w}(t)$.

MEG data is very noisy. Therefore, classifying single word waveforms yields a low accuracy, peaking at 60%, which might lead to false negatives when looking for correspondences between neural network features and brain data. To reveal informative features, one can boost signal by either having several repetitions of the stimuli in the experiment and then averaging [117] or by combining the words into larger chunks [127]. We chose the latter because the former sacrifices word and feature diversity.

At testing, we divide the consecutive words into non-overlapping segments of 20 words. For every data segment, we repeat the following: we use as possible labels the real word segment and an incorrect word segment, for all incorrect word segments (this results in 600 comparisons per fold). Since every fold of the data was used 9 times in the training phase and once in the testing phase, and since we use a high number of comparisons, this averages out biases in the accuracy estimation. Classifying sets of 20 words improves the classification accuracy greatly while lowering its variance and makes it dissociable from chance performance.

After averaging out the results of multiple folds, we end up with average accuracies that reveal how related one of the models' constituents (e.g. the RNNLM context vector) is to brain data.

Classification

In order to align the brain processes and the different constituents of the different models, we use a classification task. The task is to classify the word a subject is reading out of two possible choices from its MEG recording. The classifier uses one IFS in an intermediate classification step. For example, the classifier learns to predict the MEG activity for any setting of the RNNLM context vector. Given an unseen MEG recording for an unknown word t and two possible story words t' and t'' (one of which being the true word t), the classifier predicts the MEG activity when reading t' and t'' from their context vectors. It then assigns the label t' or t'' to the word recording t depending on which prediction is the closest to the recording. The following are the detailed steps of this complex classification task. However, as for the previous experiments detailed in this thesis, the most useful point to keep in mind is that the main purpose of the classification is to find a correspondence between the brain data and a given IFS j.

1. Normalize the columns of $\mathbf{M}$ (i.e. make $\mathbf{M}$ have a mean of 0 and a standard deviation of 1), where $\mathbf{M}$ is the data matrix, with each row corresponding to the MEG image of one word[8]. Pick feature set $\mathbf{F}_j$ and normalize its columns to a minimum of 0 and a maximum of 1.

[8]Remember for every word we have a $p = 306 \times 20$ dimensional image: we have 306 sensor locations, and for each we have 20 time point, each corresponding to the average of 25ms non-overlapping bins between 0 and 500 ms after word onset.

2. Divide the data into 10 folds, for each fold b:
   (a) Isolate $\mathbf{M}^b$ and $\mathbf{F}_j^b$ as test data. The remainder $\mathbf{M}^{-b}$ and $\mathbf{F}_j^{-b}$ will be used for training[9].
   (b) Subtract the mean of the columns of $\mathbf{M}^{-b}$ from $\mathbf{M}^b$ and $\mathbf{M}^{-b}$ and the mean of the columns of $\mathbf{F}_j^{-b}$ from $\mathbf{F}_j^b$ and $\mathbf{F}_j^{-b}$
   (c) Use ridge regression to estimate the parameters of the predictive model:
   $\hat{\mathbf{B}} = \text{argmin}_{\mathbf{B}} \ \|\mathbf{M}^{-b} - \mathbf{F}_j^{-b} \times \mathbf{B}\|_2^2 + \lambda \|\mathbf{B}\|_2^2$.
   by tuning the λ penalty to every one of the p output dimensions independently. λ is chosen via generalized cross validation [36].
   (d) Perform a binary classification. Divide the words in b into non-overlapping sets of 20 consecutive words. For every pair of sets c and d:
      i. predict the MEG data for c and d as: $\mathbf{P}^c = \mathbf{F}_j^c \times \mathbf{\Gamma}_j^b$ and $\mathbf{P}^d = \mathbf{F}_j^d \times \mathbf{\Gamma}_j^b$
      ii. assign to $\mathbf{M}^c$ the label c or d depending on which of $\mathbf{P}^c$ or $\mathbf{P}^d$ is closest (Euclidean distance).
      iii. assign to $\mathbf{M}^d$ the label c or d depending on which of $\mathbf{P}^c$ or $\mathbf{P}^d$ is closest (Euclidean distance).
3. Compute the average accuracy.

Concatenation of subject's data

Keeping with the methods we outlined in section 3.4.2, we concatenate the data from multiple subjects in order to take advantage of the fact that, while we don't have any repetitions, we do have independent samples of 8 brains reading the text with the same timings. This concatenation doesn't affect the training step since the weights for every dimension are obtained separately. Just as outlined in section 3.4.2, the combination step happens when the distance is computed between the two predictions and the real data frame, and it aims to reduce the classification error.

Restricting the analysis spatially: a searchlight equivalent

We adapt the searchlight method [65] to MEG. The searchlight is a discovery procedure used in fMRI in which a cube is slid over the brain and an analysis is performed in each location separately. It allows to find regions in the brain where a specific phenomenon is occurring. In the MEG sensor space, for every one of the 102 sensor locations ℓ, we assign a group of sensors g_ℓ. For every location ℓ, we identify the locations that immediately surround it in any direction (Anterior, Right Anterior, Right etc...) when looking at the 2D flat representation of the location of the sensors in the MEG helmet (see Fig. 5.5 for an illustration of the 2D helmet). g_ℓ therefore contains the 3 sensors at location ℓ and at the neighboring locations. While we have tried using a varying radius of locations, we report in this chapter the results of including in each g_ℓ only the three sensors at location ℓ. While this reduces accuracy, it however improves localization of the (already spatially smooth) MEG data.

[9]The rows from $\mathbf{M}^{-b}$ and $\mathbf{F}_j^{-b}$ that correspond to the five words before or after the test set are ignored in order to make the test set independent.

We also use a concatenated searchlight approach in this chapter, combining the data in one region from all the subjects. Similarly to fMRI, the subjects MEG sensor recordings do not align neatly because of the difference in the shape and size of the subjects brain. To address this, we can consider, just as we did for fMRI, using a radius for the searchlight that is wider than one location, i.e. to include more than three sensors. However, since the MEG recording is spatially smooth as mentioned above, it seems from our results that one could combine multiple subjects brains spatially using only one location at a time, i.e. it seems the spatial smoothness make MEG more forgiving than fMRI in terms of subject spatial alignment.

Restricting the analysis temporally

In addition to using the entire time course of the word, we also use each of the corresponding 25ms time windows separately. Obtaining a high classification accuracy using one of the time windows and feature set j means that the analogous type of information is encoded at that time.

Classification accuracy by time and region

Steps 1-3 above compute whole brain accuracy using all the time series. In order to perform a more precise spatio-temporal analysis, one can use only one time window m and one set of locations ℓ for the classification. This can answer the question of when and where different information is represented by brain activity. For every location, we will use only the columns corresponding to the time point m for the sensors belonging to the group ℓ. Step (d) of the classification procedure is changed as such:

(d) Perform a binary classification. Divide the words in b into non-overlapping sets of 20 consecutive words. For every pair of sets c and d, and for every setting of $\{m, \ell\}$:

   i. predict the MEG data for c and d as: $\mathbf{P}^c_{\{m,\ell\}} = \mathbf{F}^c_j \times \mathbf{\Gamma}^b_{j,\{m,\ell\}}$ and $\mathbf{P}^d_{\{m,\ell\}} = \mathbf{F}^d_j \times \mathbf{\Gamma}^b_{j,\{m,\ell\}}$

   ii. assign to $\mathbf{M}^c_{\{m,\ell\}}$ the label c or d depending on which of $\mathbf{P}^c_{\{m,\ell\}}$ or $\mathbf{P}^d_{\{m,\ell\}}$ is closest (Euclidean distance).

   iii. assign to $\mathbf{M}^d_{\{m,\ell\}}$ the label c or d depending on which of $\mathbf{P}^c_{\{m,\ell\}}$ or $\mathbf{P}^d_{\{m,\ell\}}$ is closest (Euclidean distance).

Statistical significance testing

One way to determine the distribution of classification accuracy under the null is to run a shift test similar to the one described in chapters 3 and 4. After the p-value for each accuracy value is computed, correction for multiple comparisons should be performed because the classification is repeated at every time and location $\{m, \ell\}$. Because of the strong spatio-temporal smoothness in the data (nearby points in space and time being close together), the correction would benefit from taking into account space and time to avoid grainy spatio-temporal maps. This can be achieved using cluster permutation testing [58], where the size of the largest cluster under chance performance is determined empirically, and clusters that are larger than the α quantile are rejected. However, this only allows us to reject the null hypothesis in a "weak" manner, i.e. on a cluster

basis, it doesn't allow us to state specific hypotheses about the individual sensors for which the null hypothesis is rejected (this requires "strong" control of multiple hypotheses), see [58]. We will address this in future work, however, for the time being, we present the un-thresholded results. These continuous gradings have the advantage of being more interpretable. The reader is asked to consider these results as a first approach, they are not intended to be conclusive.

5.4 Results

We present in Fig. 5.3 the accuracy using all the time windows and sensors. In Fig. 5.4 we present the classification accuracy when running the classification at every time window exclusively. In Fig. 5.5 we present the accuracy when running the classification using different time windows and groups of sensors centered at every one of the 102 locations.

It is important to lay down some conventions to understand the complex results in these plots. To recap, we are trying to find parallels between model constituents and brain processes. We use:

- a subset of the data $\{m, \ell\}$, corresponding to time window m and the group ℓ of sensor locations (for example the time window 100-125ms and all the sensors).
- one type of feature j (for example the context layer)

and we obtain a classification accuracy $a^j_{m\ell}$. If $a^j_{m\ell}$ is low, there is probably no relationship between the feature set and the subset of data. If A is high, it hints to an association between the subset of data and the mental process analogous to the feature set. For example, when using all the sensors and time window 100-125ms, along with the context layer, we obtain an accuracy of 0.92 (see Fig. 5.4). Since the context layer summarizes the context of the story before seeing word t, this suggests that the brain is still processing the context of word t between 100-125ms.

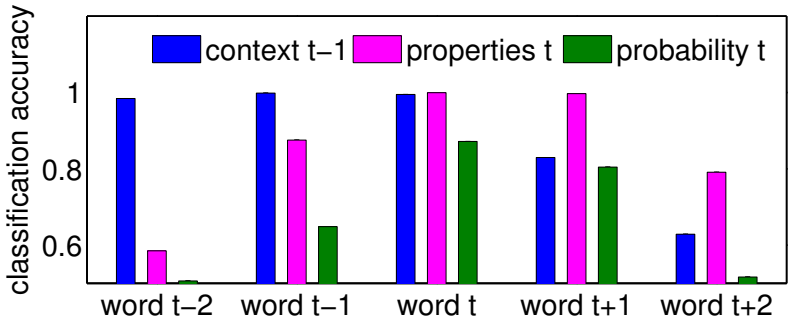


Figure 5.3: Classification for different feature types, using the entire word brain image (all time points and all sensors). The context vector at time t, the properties of word t and the probability of word t are used to predict at classify the activity for word t, but also the previous words $t-2$ and $t-1$, and the subsequent words $t+1$ and $t+2$.

Fig. 5.3 shows the accuracy for different types of features when using all time points and sensors to classify a word. We perform the classification of word t not only using the brain image of word t, but also using separately the brain images of words $t-2$, $t-1$, $t+1$ and $t+2$. The context layer features are the most powerful for classification until the word t appears. When the word t is on the screen, the context features are very good for decoding (almost 100% accuracy),

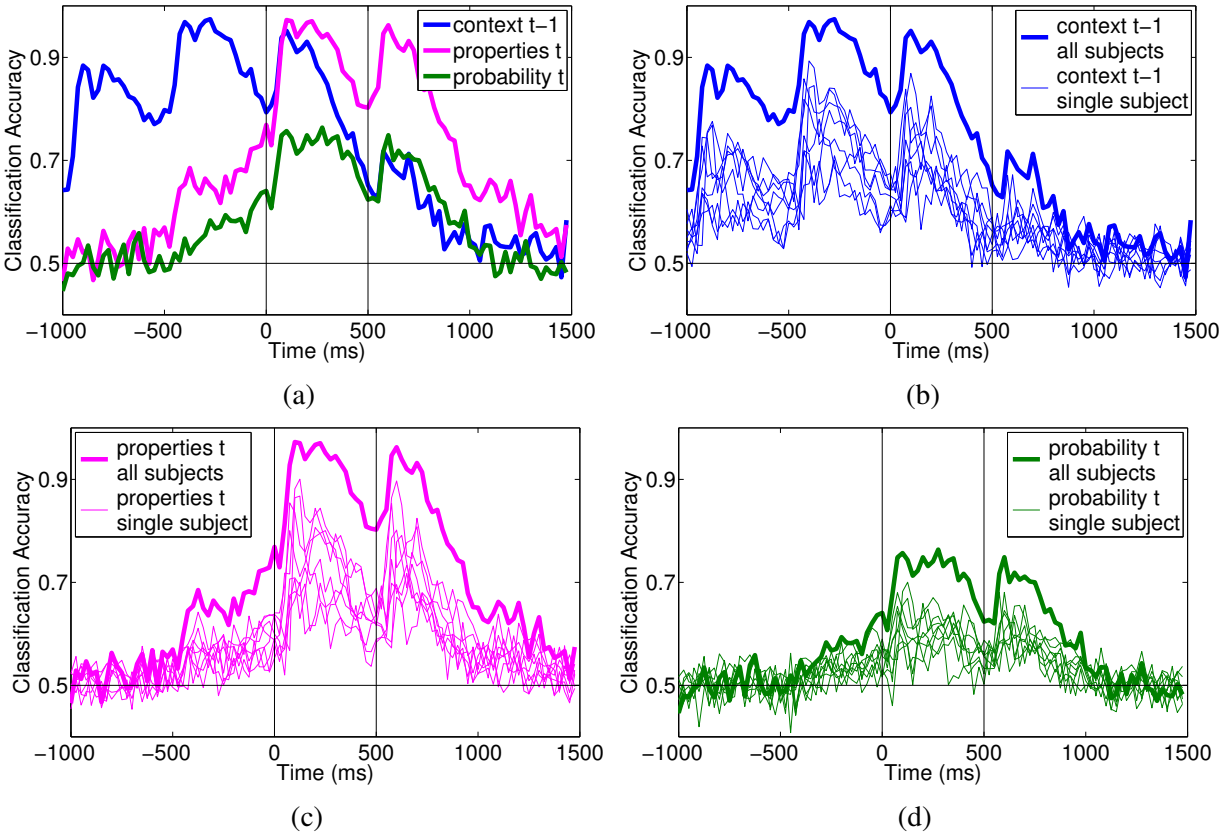


Figure 5.4: Classification accuracy over all sensors by time window when using the context vector, the properties vector and the probability of word t. Panel 5.4a shows the accuracy when combining all subjects, while panels 5.4b to 5.4d show the accuracy both for the combination of subjects and for individual subjects, for the three feature types. Word t's onset (time 0) and offset (time 500ms) are marked with vertical lines.

and seemingly just as good as word t's properties at capturing the information contained in the brain data, suggesting that a lot of the brain activity is encoding the previous context.

The properties of the word t lead to high accuracy before and after the word is on the screen, which might have multiple explanations. There might be correlations between the properties of consecutive words, causing the property vector at time t to predict the activity of surrounding words. The brain might be predicting the properties of the next word at time $t-1$, causing the brain activity at time $t-1$ to be correlated with the properties of word t, and therefore driving the accuracy up. A third hypothesis is that the brain keeps processing the properties of word t when word $t+1$ is on the screen, which makes the properties at time t able to well predict the activity at time $t+1$. We address these three hypotheses in the next part of this chapter.

Finally, output probability has the smallest accuracies. This makes sense considering that it captures much less information than the other two high dimensional descriptive vectors, as it does not represent the complex properties of the words, only a numerical assessment of their likelihood. Unexpectedly, the output probability of word t is able to predict the activity of words $t-1$ and $t+1$. Since intuitively unexpected words can only affect brain activity after they occur,

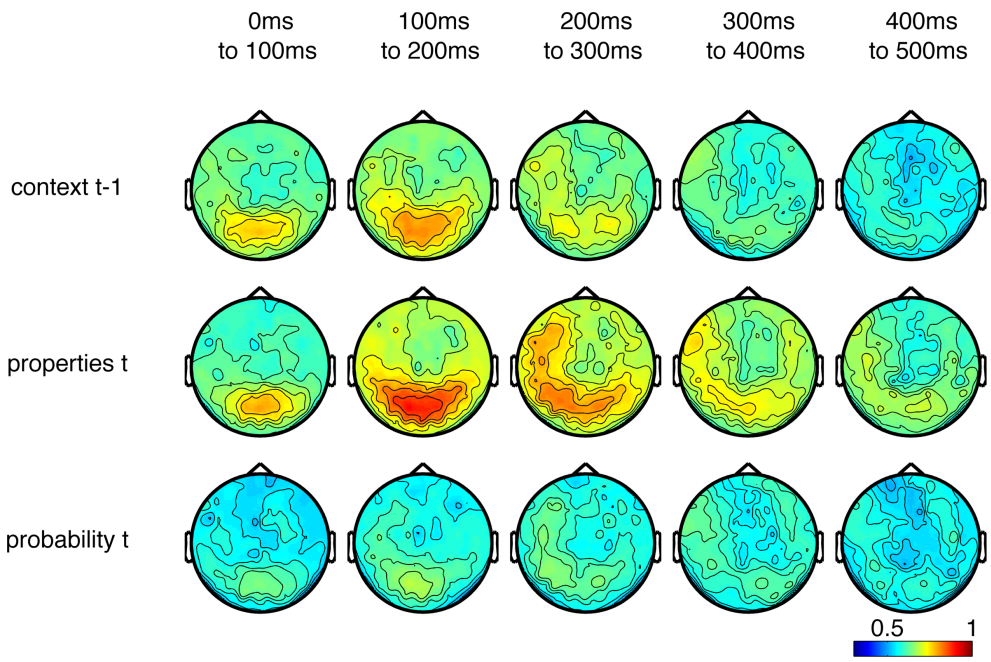


Figure 5.5: Classification accuracy by sensor and time window when using the context vector, the properties vector and the probability of word t. The accuracy, originally computed for each 25ms window, is averaged into bins of 100ms for space concerns. The contour plots are obtained by interpolating the accuracy across the sensor locations.

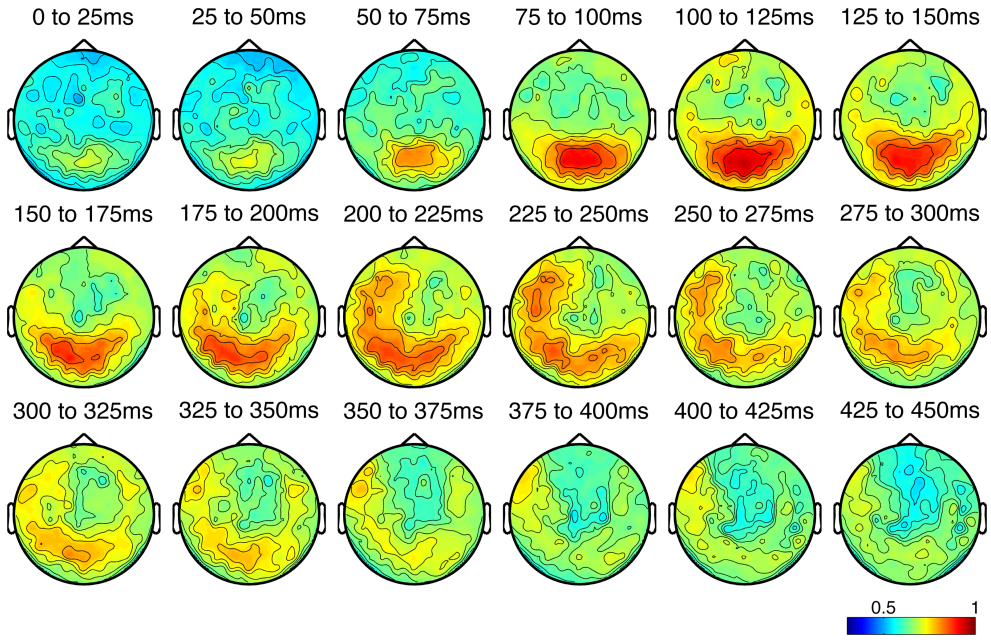


Figure 5.6: Classification accuracy by sensor and time window when using the **properties vector** of word t. We provide here all the time points so the reader can see the gradual progression of the decoded representation from the posterior visual regions (around 100ms), to the posterior left temporal cortex (around 200 ms), to more anterior parts of the left temporal cortex (around 250).

it seems likely that the probability vector is confounded with some other variables.

Fig. 5.4a shows the accuracy when using different windows of time exclusively, for the 25ms time windows starting at $-1000, -975 \ldots 1475$ms (where 0 is the onset of word t). To understand the time dynamics of the context vector accuracy we need to see a larger time scale than the word itself. The context vector captures the context before word t is seen. Therefore it seems reasonable that the context vector is not only related to the activity when the word is on the screen, but also related to the activity before the word is presented, which is the time when the brain is integrating the previous words to build that context. On the other hand, as the word t and subsequent words are integrated, the context starts diverging from the context of word t (computed before seeing word t). We see the behavior we predicted in the results: the context before seeing word t becomes gradually more useful for classification until word t is seen, and then it gradually decreases until it is no longer useful since the context has changed.

The accuracy of the properties of word t start increasing before the word t is on the screen. As hypothesized above, this might be due to the brain predicting word t at time $t-1$, or simply to correlations between the properties of word t and the properties of word $t-1$. The properties of word t are able to predict the signal well after word t is off the screen. Interestingly, the properties waveform at times 0-1000ms are very similar to the context waveform at times -500-500ms, which hints strongly that the context vector is highly related to the properties of word $t-1$ (we know that the context vector is obtained by combing the properties of word $t-1$ with the context vector at $t-2$). The next section will address this issue. Finally, the probability vector predicts brain activity before and after the word is on the screen, highlighting the need for a deeper approach to figure out what is happening under the hood.

To show the consistency of our results, we show in panels 5.4b-5.4d the accuracy in time for the combination of subjects and per subject, for the three feature sets. The patterns seem very consistent, indicating the phenomena we described is robust and can be detected at the subject level. This similarity of results is remarkable for (1) MEG and (2) an experiment with zero repetitions. The second interesting point in panels 5.4b-5.4d is the considerable improvement we obtain by concatenating the subjects.

We now move on to the spatial decomposition of the analysis. When the visual input enters the brain, it first reaches the visual cortex at the back of the head, and then moves anteriorly towards the left and right temporal cortices and eventually the frontal cortex. As it flows through these areas, it is processed to higher levels of interpretations. In Fig. 5.5, we plot the accuracy for different regions of the brain and different time windows for the three feature sets. There is a similarity between the patterns for the three feature sets that invites us to investigate further the possibility of spurious correlations, and we perform this analysis in the next section.

In Fig. 5.6 we plot the progression of the accuracy for the properties of word t in detail, for every 25ms. We see that in the back of the head, in the visual cortex, the word properties have an accuracy that peaks very early on (around 100ms). At around 200ms the activity in the posterior temporal cortex starts being related to the properties vector, reflecting the delay it takes for the information to reach this part of the brain. This relationship moves even more anteriorly and by 250ms it reaches the anterior temporal cortex. Because of its ability to correctly predict the signal in such a long stretch of the brain, seemingly highlighting the entire process of word perception after word onset, the properties vector seems to encompass a lot of the features of the words at different levels of abstraction (e.g. syntax, semantics etc.). We analyze this in section

5.6. In the next section, we will see that this pattern successfully survives our attempt to control for correlations.

5.5 A classification test for conditional independence

Our previous decoding experiment showed a very strong relationship between the properties of word t and the brain activity during word $t+1$. We are interested in finding out whether this activity is due to the brain still processing word t at time $t+1$, or if this is due to some spurious correlations between the properties of consecutive words. Does the brain process word t and word $t+1$ simultaneously? Are these processed in the same regions?

In order to answer that question, we need a conditional independence test. We need to find if the features of word t still explain brain activity even after we take into account the contribution of word $t+1$ to the brain activity, and vice versa. As we discussed in chapter 6, conditional independence tests are hard in settings with continuous, multivariate variables, and we are currently working on adapting existing methods to the complex setting of IFS analysis. We propose to run the following test as a measure of conditional independence. Given brain activity $\mathbf{v}_{m\ell}$ in region ℓ at time offset m and feature sets $\mathbf{f}_1, \mathbf{f}_2, ...\mathbf{f}_F$, we test if $\mathbf{f}_j$ is related to $\mathbf{v}_{m\ell}$ given $\{\mathbf{f}_i\}_{i \neq j}$ in the following way:

1. Perform the IFS classification as outlined above with $\mathbf{v}_{m\ell}$ and $\mathbf{s} = [\mathbf{f}_1, \mathbf{f}_2, ...\mathbf{f}_F]$ to obtain the accuracy $a_{m\ell}$ ($\mathbf{s}$ is the concatenation of all the feature vectors of interest).
2. Perform the IFS classification as outlined above with $\mathbf{v}_{m\ell}$ and $\mathbf{s}^{-j} = [\mathbf{f}_1, ...\mathbf{f}_{j-1}, \mathbf{f}_{j+1}, ...\mathbf{f}_F]$ to obtain the accuracy $a^{-j}_{m\ell}$ ($\mathbf{s}^{-j}$ is the concatenation of all the feature vectors of interest except for $\mathbf{f}_j$).
3. Use $c^j_{m\ell} = a_{m\ell} - a^{-j}_{m\ell}$, the improvement in accuracy due to feature $\mathbf{f}_j$, as a statistic to test if $\mathbf{f}_j$ is related to $\mathbf{v}_{m\ell}$ given the other feature sets. Test for significance appropriately.

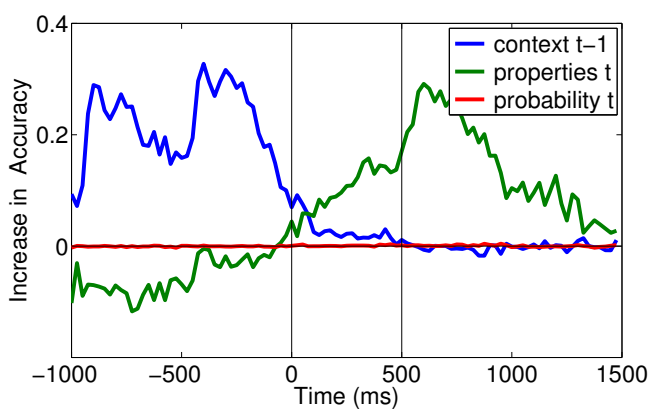


Figure 5.7: Increase in classification accuracy due to the inclusion of a feature set (context, properties or probabilities). The plots show c^j_m, the increase in accuracy per time window between performing the classification with (1) context, properties and probabilities and (2) with only two of these sets with the third omitted (indicated by color). This statistic is a tool to assess if if there is a part of MEG activity that is related to feature j that is not accounted for by the other features.

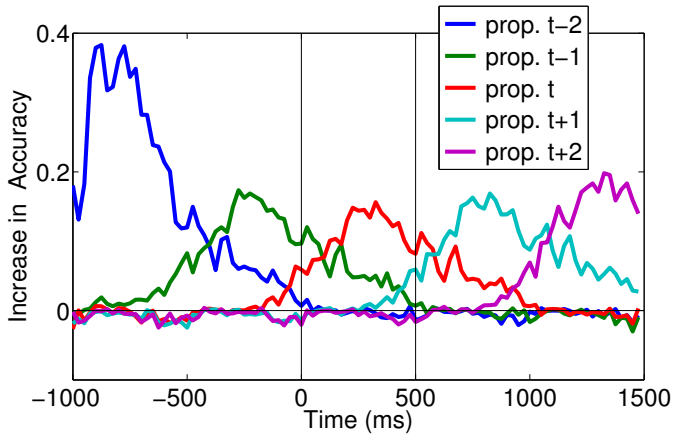


Figure 5.8: Increase in classification accuracy due to the inclusion of a feature set (properties of word $t-2$, $t-1$, t, $t+1$ and $t+2$). The plots show c_m^j, the increase in accuracy per time window between performing the classification with (1) the set of properties of all words $t-2$ to $t+2$ and (2) the set of properties of words $t-2$ to $t+2$ **excluding** the properties for one of these words (indicated by plot color).

5.5.1 Distinction between current word properties and context

In figure 5.7 we show the results of the classification test. We plot $c_m^{\text{CONTEXT (t-1)}}$, $c_m^{\text{PROP. (t)}}$ and $c_m^{\text{PROB. (t)}}$, i.e. respectively the increase in accuracy when adding the context at $t-1$, the properties of t or the probability of t (to the other two features sets). We see that during word t, the properties of word t can improve upon the accuracy using the context $t-1$ and the probability of t. $c_m^{\text{PROP. (t)}}$, during the time when word $t+1$ is on the screen, seems like it has not been appropriately "controlled": it is even higher than when the word t itself is on the screen. In other words, it seems like the context vector at time $t-1$ predicts a considerable portion of the MEG data at time t, which the properties of t vector also predicts, however, during word $t+1$, there is a decrease in the ability of the context vector to predict the activity. This seems to be why the properties t can improve the accuracy to a greater extent during $t+1$ than during t. In order to account for this, we will now run the analysis while explicitly controlling for the features of words t, $t+1$ and $t+2$, as well as the previous words $t-2$ and $t-1$.

5.5.2 Word processing even after the next word appears

In figure 5.8 we show the improvement in classification accuracy when the property vector at time t' (such that $-2 \leq t' \leq 2$) is added to the set $\{t-2, t-1, t, t+1, t+2\} - \{t'\}$, i.e. $c_m^{\text{PROP. (t')}}$. We can see each property vector vector t' is useful for classification when the word t' is on the screen, and continues on for a few hundreds of milliseconds after the next word appears. The peak for feature set $t-2$ is higher than the other words because we don't include previous words to $t-2$ that would "control" its marginal accuracy. Otherwise, the accuracies for these consecutive property vectors should be translated curves (the feature sets themselves are translations of each other). We illustrate this point as well in the spatio-temporal analysis in Fig. 5.9. This figure shows the spatio-temporal increase in activity for properties t', which also peaks when the word t' is on the screen and lasts for a few hundreds of milliseconds after the next word appears. We can also see a similar pattern to figure 5.6 where the accuracy starts increasing in the visual cortex, and then moves anteriorly, mainly in the left temporal cortex.

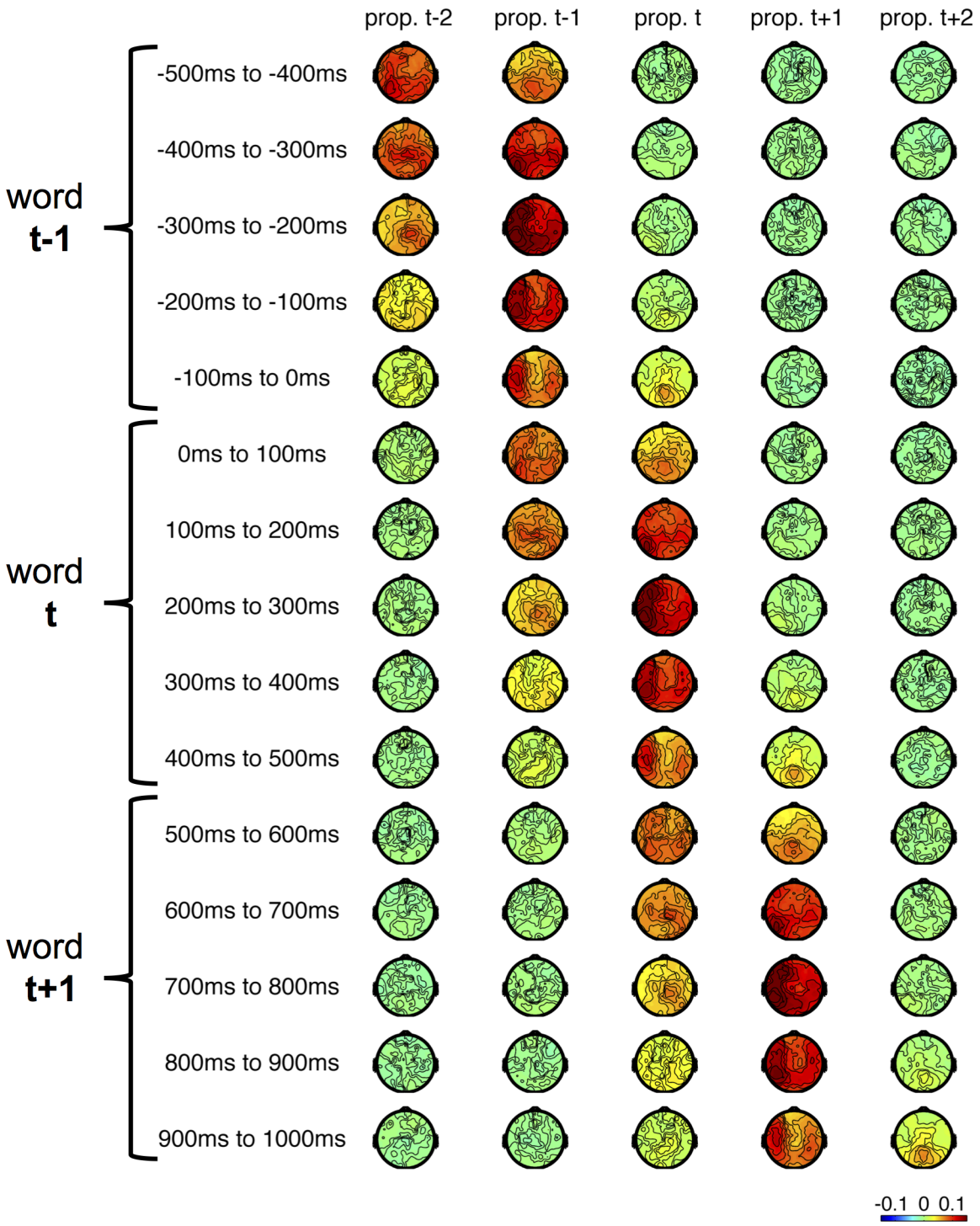


Figure 5.9: Increase in classification accuracy due to the inclusion of a feature set (properties of word $t-2$, $t-1$, t, $t+1$ and $t+2$) for different regions and time points. Word t appears at time 0. The accuracy is averaged into bins of 100ms for space concerns. The brain activity seems to be related "specifically" to the features of word t starting when word t is on the screen and lasting until 700-800ms later. The word is replaced at 500ms, which suggests the brain is still processing word t when it is perceiving word $t+1$.

We highlight the change in the spatio-temporal pattern of improved accuracy in figure 5.10, which displays the 25ms changes across the brain when using the properties of word t. The improvement is computed after adding the properties of word t to the set of properties $t-2$, $t-1$, $t+1$ and $t+2$. We can see a very similar pattern to 5.6, confirming that the properties derived from the neural networks are allowing us to uncover the progressive perception of the word starting from the visual system to regions that are more implicated in language like the left anterior temporal cortex. Again, these properties seem to be encoding a lot of the features of the words along multiple levels of processing. The next section will address this question and try to find what exact type of information each of the regions is processing.

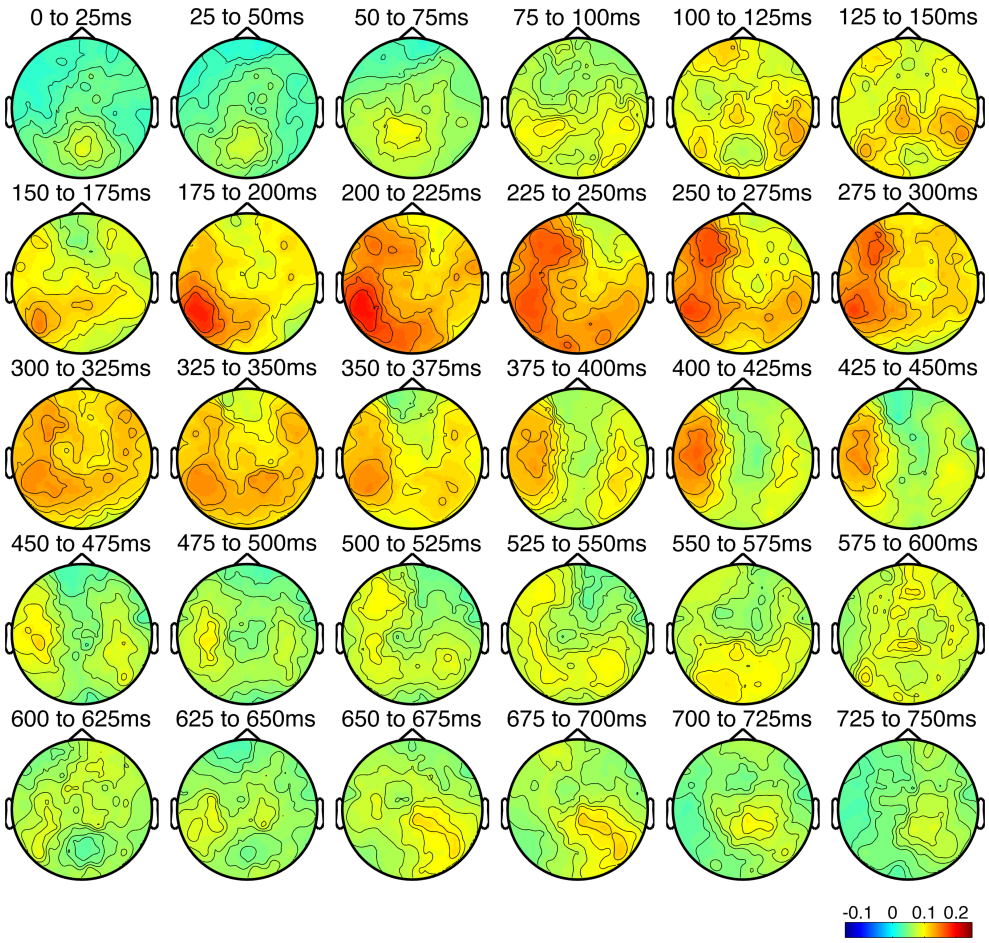


Figure 5.10: Increase in classification accuracy by sensor and time window when including the **properties vector of word t**. The increase we plot is specific to adding the properties of word t to the set of properties of words $t-2$, $t-1$, $t+1$ and $t+2$. We see a similar pattern of progression of representation than figure 5.6, which adds evidence to the general pattern we observed in that "uncontrolled" setting. The representation moves from the posterior visual regions (around 100ms), to the posterior left temporal cortex (around 200 ms), to more anterior parts of the left temporal cortex (around 250) and is processed in the left temporal cortex for a while (until around 450ms).

5.6 Accessing different word properties

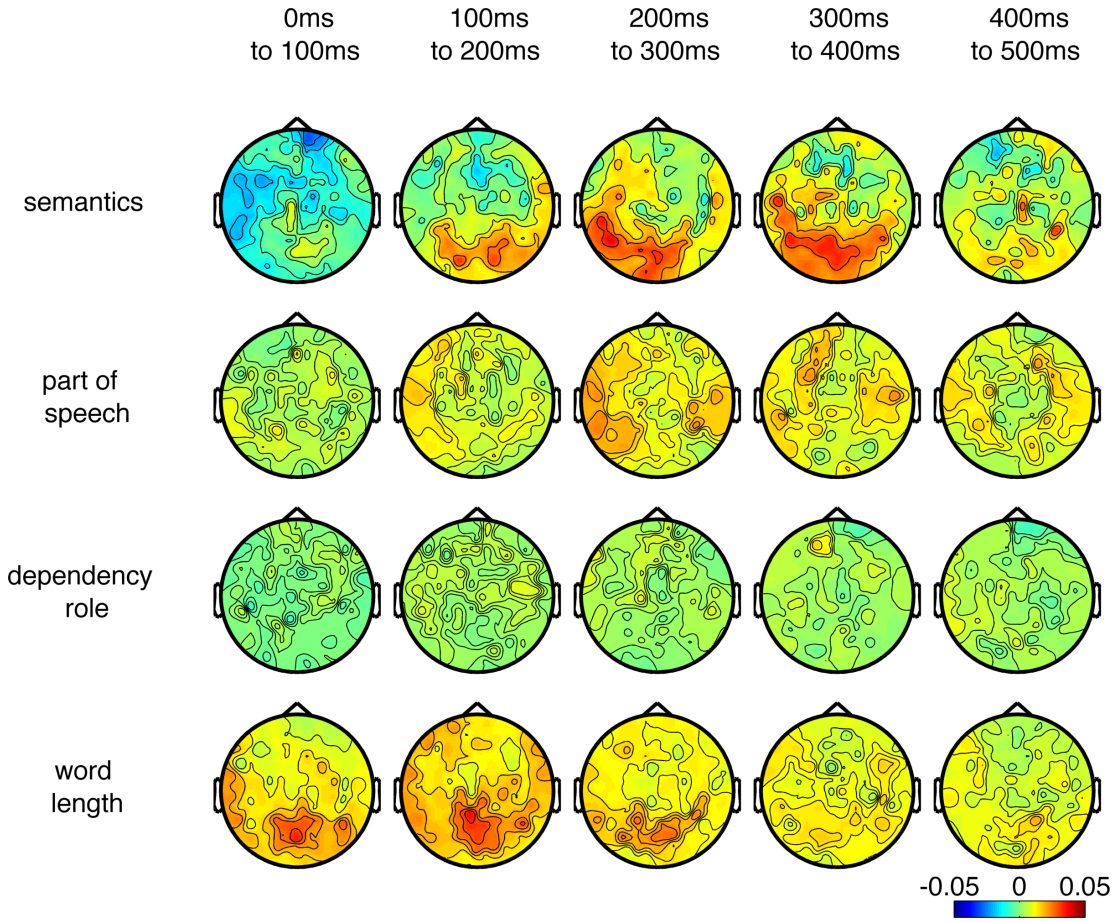


Figure 5.11: Increase in classification accuracy by sensor and time window when including a feature vector. The total set of features being considered is: **context, semantics, part of speech, dependency roles** and **word length**, i.e. the results being plotted are increases after the context of the word and three other feature sets have been accounted for.

In order to find when and where different word properties are processed by the brain, we replace the uninterpretable RNNLM properties vector with word specific features similar to the features we chose in chapter 4. Specifically, we use **in conjunction with the context vector**:

1. Corpus based (single-word) semantic vectors obtained by counting concurrence frequencies of words (within a window of size 4 on either side of the word) and reducing dimensionality using PCA to 300 dimensions. We used the features obtained by and described in [85] (we use the feature set referred to as "Word-Form").
2. The same part of speech features as in chapter 4.
3. The same dependency role features as in chapter 4 (i.e. the grammatical role of each word in its sentence).
4. The length of every word (in number of letters).

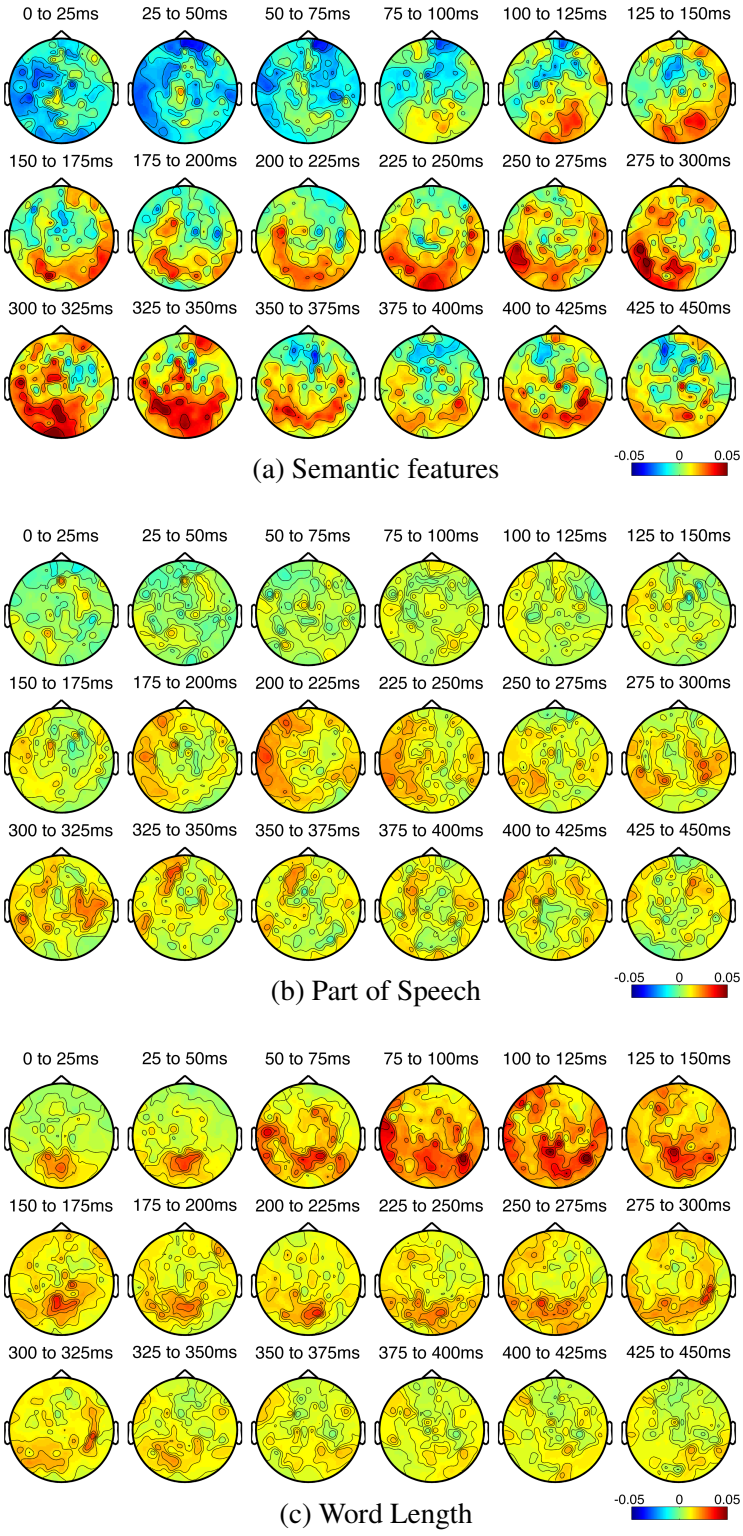


Figure 5.12: Increase in classification accuracy by sensor and time window when including a feature vector. The total set of features being considered is: **context, semantics, part of speech, dependency roles** and **word length**.

Fig. 5.11 shows the increase in accuracy in different regions and at different times when using the different types of word features. The first feature to have an improvement in accuracy is the word length feature, starting to peak before 100ms and lasting only until 300ms mostly in the visual cortex. After the visual properties peak, the semantic features and the part of speech features start improving accuracy. The semantic features initially peak very posteriorly, in what looks like the visual cortex. We are currently working a source localized version of these experiments in order to assess what region is exactly responsible for semantic representation in or around the visual cortex. The peak in increase of accuracy due to the semantic features then moves after 200ms to the posterior temporal cortex. As for part of speech, the improvement in accuracy seems more anterior than semantics, and has a small right hemisphere representation in parallel with the left. The dependency features do not seem to have a substantial improvement in accuracy (however, see the disclaimer in section 5.7).

Figures 5.12a to 5.12c show in detail the 25ms spatio-temporal progression for the semantic features, the part of speech features and the word length features, so that the interested reader can judge them more easily.

5.7 Perspective

Disclaimer

Our classification test for conditional independence aims to find regions and times where a portion of the activity is uniquely explained by feature set j. However, that does not mean that j is not expressed in the brain at other regions and times. j might be simultaneously processed with another feature set j' by a region that is responsible for the common core between j and j'. For example, let's assume that being a noun is both a semantic and a syntactic attribute. Then a region which process the fact of being a noun might be overlooked by our conditional independence test, however, it is processing "semantics" and "syntax (under the above assumption).

Based on this, we will attempt an explanation to the contradictory results in the literature about the representation of syntax and semantics. For example, in [22], the authors do not find a region in the language system that is responsive uniquely to syntax or semantics. There might that the parts of the language system are processing multiple language processes (which leads to common discrepancies in results between studies), but however, these differ in how much they are involved in each process. An expressive modeling approach such as ours might be key to solving this problem.

Novel brain data exploration

We present here a novel and revealing approach to shed light on the brain processes involved in reading. This is a departure from the classical approach of controlling for a few variables in the text (e.g. showing a sentence with an expected target word versus an unexpected one). We are able to extract from the data many more details about what is being resented, where and when, and offer a much richer interpretation than is possible with artificially constrained stimuli.

Comparing two models of language

We showed that it might be possible to use brain data to understand, interpret and illustrate what exactly is being encoded by the obscure vectors that neural networks compute, by drawing parallels between the models constituents and brain processes. This is just a start, but the field of computational neuroscience of language and the field of natural language processing which aims to automate language comprehension and production are trying to solve different facets of the same problem. We should draw example on the close collaborations done between the fields of computational neuroscience of vision and computer vision, and group the two language processing fields under the same investigation philosophy.

A research direction we are interested in is using brain data to improve statistical language models. The brain recordings of subjects performing the same task an algorithm is trying to achieve can be thought of as noisy "clues". These clues are given by a computational system that knows how to perform the task: the brain. Constraining language models with brain activity recordings of people reading *might* therefore bias the learning of these models in the correct direction.

Future work

The work described here is our first attempt along the promising endeavor of matching complex computational models of language with brain processes using brain recordings. We plan to extend our efforts by (1) building a model that jointly predict the next word and the next word's brain activity and (2) make the brain data help us with training better statistical language models by using it to determine whether the models are expressive enough or have reached a sufficient degree of convergence.

Unexpected achievements

We have been able to get notable achievements that are for the most part unprecedented. We used MEG (notorious for its low signal to noise ratio) in a single trial setting with no repetitions (which is very rarely done) in order to process a very complex, naturalistic paradigm (naturalistic tasks are also rarely done, although they are becoming more popular).

We were able to obtain very consistent results among subjects, and we were able to boost the classification accuracy considerably by applying our technique of concatenating MEG data from multiple subjects. Perhaps the most surprising is that we were able to get such a clear spatial differentiation between the different features sets, while working in sensor space[10].

[10] Admittedly, this might be due to the fact that classification independence test looks for regions where the performance of a feature set is different from others

Chapter 6

One Step Hypothesis Testing

This thesis has so thus consisted in approaches to analyze *naturalistic* neuroimaging experiments, which are presently rather uncommon. We have seen that unlike traditional *controlled* experiments which are only designed to isolate and study one process, naturalistic experiments allow us to study the interaction of various brain processes, but that they however are considerably more difficult to analyze, requiring from the analyst many choices of complex multistep methods. These estimation (regression/classification) based procedures may cause a lack of reproducibility of results; their non-robustness may be caused by biases from unmet parametric assumptions or model misspecification.

We propose in this chapter a unified statistical framework that encompasses both naturalistic experiments, posing them as *independence tests*, and controlled experiments, posing them as *two-sample tests*. For both setups, we advocate the *direct* use of single-step model-free nonparametric hypothesis tests that require much weaker assumptions than estimation based methods. With this new, simple and direct framework at hand, one can now reliably analyze naturalistic experiments to complement (but not replace) controlled experiments. Instead of the complex setup of naturalistic experiment analysis we introduced and used in the previous chapters, this framework results in a clear formulation of the problem as a simple independence test, which can be easily extended to account for temporal delays in the signal. This will allow the problem and solution to be more accessible to researchers across disciplines. We illustrate our framework by analyzing the naturalistic reading experiment from chapter 4, identifying dependencies between the text properties and the activity of brain regions at different latencies.

Brain imaging data is rich and complex, even for simpler controlled experiments. Brain images are noisy and high dimensional (fMRI images have in the order of 30,000 voxels, or "volume pixels") and have a rich spatial structure. This has made brain data appealing to machine learning researchers, causing a cross-fertilization of fields like cognitive neuroscience and applied statistics. Intricate and innovative methods have been applied to it, as evidenced by a multitude of recent papers in various fields; as an example see the recent ML conference papers [71, 103, 129]. For naturalistic experiments, the problem becomes even more complex. Naturalistic stimuli vary along a very large number of dimensions, e.g. videos can be described by a great number of visual and motion features, as well as semantic features describing their scenes. At the same time, due to the measurement's nature, and the fact that the stimulus is shown continuously, the recorded data is a time series which has rich temporal structure. Furthermore, a

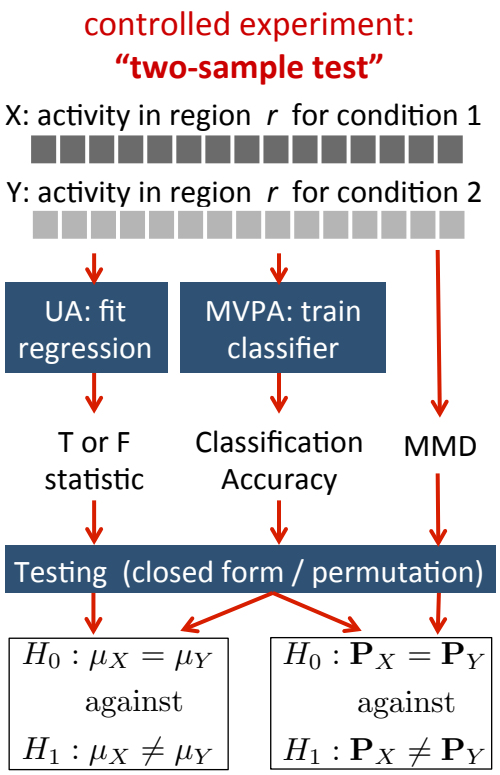
typical experiment does not last more than 1-2 hours, leading to a small number of brain images with respect to the dimensionality (in fMRI, less than 2000 samples are typically collected). For these reasons, the analysis of naturalistic brain data is an interesting problem of learning relationships in a small data setting; it promises many opportunities and challenges to the ML researcher who is interesting in coming to the rescue.

Before exploring the space of models that can be applied to this problem, caution should be exercised. Naturalistic brain imaging introduces many statistical challenges to the problem of inferring cognitive processes from brain activity because the different processes are not controlled, i.e. the discovered brain signature of one process can be confounded by spurious correlations with another. If one temporarily gives up a search for causal relationships between the input stimulus and brain activity, and the new goal is to find interesting dependencies that can be later investigated in controlled experiments, a new peril awaits. The temptation of creating new models which can explain the data well has led to a slew of different kinds of methods over the years. There are no clear guidelines which method is the best to follow. Therefore the multitude of these methods results in variability and lack of reproducibility of results, which sometimes causes negative publicity in popular press and concern in academic circles. We suggest to formulate the problem in a framework that would (1) minimize the number of model assumptions and parameter choices that are required, or even avoid a model entirely and (2) express the problem in simple statistical terms that would pave the way for future work to account for confounding variables even when the experiment is not controlled.

We present in this chapter a unified, simple statistical framework for studying both naturalistic and controlled experiments. Firstly, this framework encompasses techniques that are most commonly used to analyze controlled experiments, and are described in statistical literature as *two-sample tests*. Once this (maybe obvious in hindsight) observation is made, it allows us to advocate the use of a nonparametric kernel test from the recent ML literature. This test makes no assumptions on the distribution of the signal variables in consideration, as well as no assumptions about the noise, about a generative model or about how signal variables may differ, etc. Secondly, we frame the problem of analyzing naturalistic experiments as an *independence test*, and once again this observation allows us to advocate the use of a nonparametric kernel test. This test requires no modeling or distributional assumptions, or assumptions on the nature of the relationship between variables, etc. Furthermore, we use a shift-test, allowing us to easily account for the natural lags in brain responses relative to stimulus presentation, thus completing our framework for analyzing naturalistic experiments.

Not every single use of controlled and naturalistic experiments falls under these categories. For example, in controlled experiments one may want to *estimate* (not just test) the difference in brain activity in the two settings; for naturalistic settings, one may want to estimate (not just test) the strength of a relationship between stimulus and activity. Nevertheless, ours is a useful abstraction that encapsulates a variety of existing controlled experiments, and more importantly, it provides an intuitive way to handle and understand the complexity of naturalistic experiments. This is the first attempt at unifying how we understand controlled and naturalistic experiments under a single, statistically well-studied, umbrella and suggesting a framework for analyzing naturalistic experiments.

6.1 Two-sample testing



In this chapter, we focus on fMRI experiments, although the examples can be applied to other functional neuroimaging tools as well, such as Magnetoencephalography (MEG). We will use a capital, non-bold letter to refer to a multivariate random variable (e.g. X), and, in keeping with the notation in this thesis, a lowercase bold letter to refer to an instantiation of this variable (e.g. $\mathbf{x}_i$) and a capital bold letter to refer to the set of all instantiations of this variable during one experiment (e.g. $\mathbf{X}$ is the set of $\mathbf{x}_i$s).

The most common method for analyzing fMRI data is classical univariate analysis (UA). The aim is to find regions of the brain that represent a condition of interest, by analyzing the activity of each voxel independently. Recently, multi-voxel pattern analysis (MVPA) has emerged as an alternative. MVPA looks at the signal over brain regions, and can therefore detect brain representations that manifest as subtle differences over a group of voxels, which might be overlooked by UA.

In UA, the activity of a single voxel is modeled as a function of the few experiment conditions[1], to find if voxel responds to a condition **A** significantly more than rest, or if it responds to **A** significantly more than another condition **B**. This is tested by checking if the regression parameters for each condition are significantly different from zero (using a T-statistic) or significantly different from each other (using an F-statistic).

Brain decoding [80], or MVPA [93], consists of training a classifier to distinguish the brain state of a the subject (e.g. whether they see a face or a house) from their brain image. In other words, the decoder takes as input the brain activity, and is able to predict the label of the corresponding stimulus. Many classifiers differing in complexity have been used to distinguish between a large variety of conditions [71, 103].

Both UA and MVPA are effectively performing, in statistical terminology, **two-sample tests**, even if they are not always stated this way. Given sets of samples from two distributions ($\mathbf{x}_1, ..., \mathbf{x}_m \sim \mathbf{P}_X$ and $\mathbf{y}_1, ..., \mathbf{y}_n \sim \mathbf{P}_Y$), a two-sample test tries to detect if the two distributions are different ($\mathbf{P}_X \neq \mathbf{P}_Y$). In our setup, the two sample test consist of finding if a voxel r (or brain region r) processes condition **A** differently from **B**. In other words, under this test, the null hypothesis is that the fMRI responses in r have the same distribution under conditions **A** and **B**, i.e. r doesn't respond systematically differently for the two conditions. The alternative hypothesis is that the fMRI responses have different distributions under **A** and **B**.

Both UA and MVPA perform the two-sample test in an indirect fashion that requires the

[1]Typically each voxel's activity is regressed on a design matrix indicating the occurrence of the conditions.

intermediate estimation of a model. In UA, the statistic of interest is usually a T- or F-statistic, and the p-value can be estimated parametrically in closed form. In MVPA, the statistic is usually the average classification accuracy, and the p-value is estimated by a permutation test (discussed below).

When the ultimate goal is to perform a two-sample test, these indirect measures - that require the assumption of a model class and the selection of a model - are unnecessary. Our first argument against these indirect hypothesis tests is the intuition that they are solving a harder problem than what is required: estimating a vector (i.e. in regression or classification) versus testing (which has a single binary output). The estimation problem might constitute a bottleneck to the simpler detection problem. Other practical arguments can be made. Training and testing a classifier requires several parameter and setup choices, and it is not clear which have the most power. Fitting a regression in UA also relies on a model assumption (usually linearity). UA is also restricted by being univariate.

We therefore need a test that (1) does not require the intermediate estimation of a model, and (2) makes no assumptions on the underlying distributions of X and Y (i.e. it would be able to detect differences between these distribution without being constrained to differences in specific parameters like the mean). We need this test to (3) be multivariate so it can be applied directly to our high dimensional brain vectors, without having to estimate a model to produce a single statistic and then submit it for testing. The Maximum-Mean Discrepancy (MMD) is a test statistic that has been developed precisely to have these properties [41]. It is non-parametric, does not need to estimate an intermediate model or assume an underlying distribution, and is adaptive to non-linear settings where the distributions differ in higher moments than the mean. Additionally, it is multivariate. The MMD measures the distance between the mean functions of $\mathbf{P}_X$ and $\mathbf{P}_Y$ in a reproducing kernel Hilbert space (for more details, the reader is asked to refer to [41]). The MMD is used to test:

$$H_0 : \mathbf{P}_X = \mathbf{P}_Y \quad \text{against} \quad H_1 : \mathbf{P}_X \neq \mathbf{P}_Y.$$

This statistic is defined such that the population $\text{MMD}^2[X, Y] = 0$ if and only if $\mathbf{P}_X = \mathbf{P}_Y$. In practice, the following empirical estimator is used for m samples $\mathbf{x}_i$ and n samples $\mathbf{y}_j$:

$$\text{MMD}_u^2[X, Y] = \frac{1}{m(m-1)} \sum_{i \neq j = 1}^{m} k(\mathbf{x}_i, \mathbf{x}_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(\mathbf{x}_i, \mathbf{y}_j) + \frac{1}{n(n-1)} \sum_{i \neq j = 1}^{n} k(\mathbf{y}_i, \mathbf{y}_j)$$

We consider here the Gaussian kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp(-||\mathbf{x}_i - \mathbf{x}_j||^2)/2\sigma^2$ (any characteristic kernel can be chosen, for details see [43]). A typical method of choosing the bandwidth $2\sigma^2$ is the median heuristic [113]. $2\sigma^2$ is set to the median of the square distances $||\mathbf{z}_i - \mathbf{z}_j||^2$ between different samples $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$. This has been shown to work well for testing, however, the Gaussian MMD test is *consistent* for any bandwidth choice [43], i.e. given enough samples from $\mathbf{P}_X$ and $\mathbf{P}_Y$, this test will be able to detect any difference between $\mathbf{P}_X$ and $\mathbf{P}_Y$. A linear two-sample test can differentiate between $\mathbf{P}_X$ and $\mathbf{P}_Y$ when they differ in their means. However, the Gaussian-MMD can differentiate between $\mathbf{P}_X$ and $\mathbf{P}_Y$ when they differ in their means, variances or higher order moments.

We declare the null to be true whenever $\text{MMD}_u^2[X, Y]$ is "close" to zero in finite samples, smaller than some threshold. A permutation test is used to empirically estimate the threshold by

simulating the null hypothesis. The samples of X and Y are permuted: the samples $\mathbf{Z} = \mathbf{X} \cup \mathbf{Y}$ are randomly assigned to two sets of size m and n. This effectively simulating the case where the X and Y samples are interchangeable, i.e. they are drawn from the same distributions. The permutation is repeated a large number of times and MMD_u^2 is computed for each time. These permuted statistics are used to estimate the p-value for the real, unpermuted MMD_u^2, i.e. we compute the fraction of times a value at least this extreme is obtained with the permuted samples. The permutation test is known to be exact, but because one never goes through all permutations, a small approximation error is introduced (which is typically small enough if one does about 500-1000 permutations).

It is unclear what the power of classification accuracy is when used as a two-sample test statistic, even for kernelized classifiers like kernel-SVM. Since theoretical advancement is still far from solving this problem, we used a simulation experiment to compare the power of (Gaussian) kernel-SVM and Gaussian-MMD at distinguishing between $\mathbf{P}_X$ and $\mathbf{P}_Y$ when they are sampled from two different multivariate Gaussian distributions (Fig. 6.1). We vary the sample size, the dimensionality, and the extent to which the distributions differ. In all the experiments we tried, we see that MMD never performs worse than SVM and very often beats it.

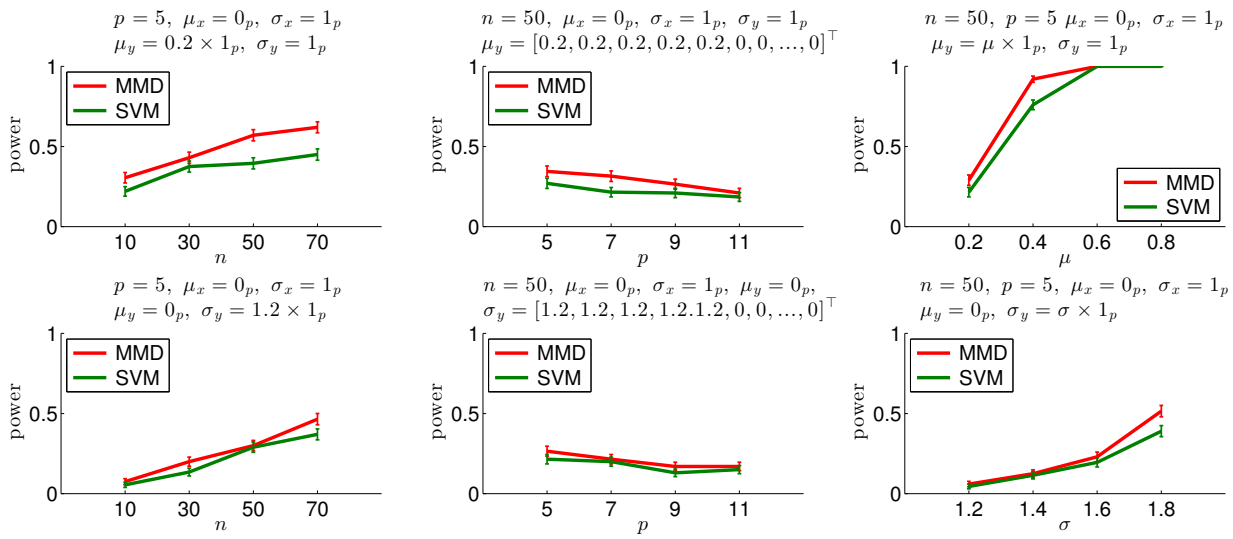
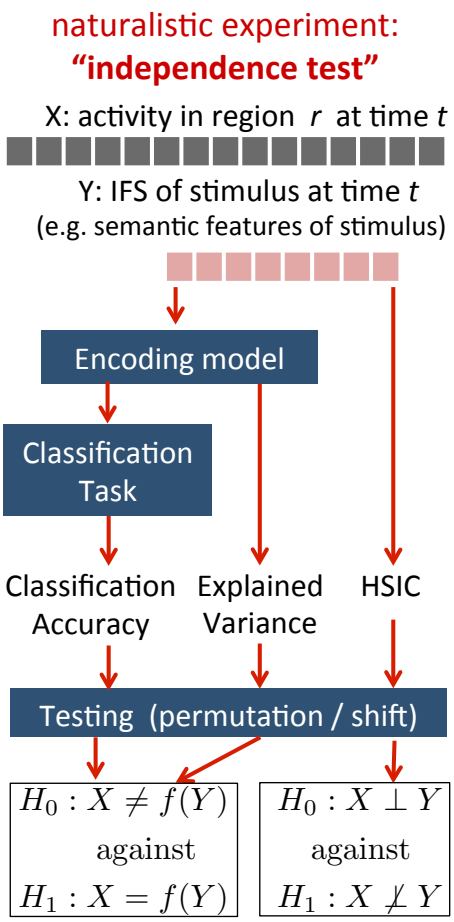


Figure 6.1: Power of MMD and an RBF SVM classifier at detecting the alternate hypothesis of $\mathbf{P_x} \neq \mathbf{P_y}$. We simulate $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^p$ such that $\mathbf{P_x} = \mathcal{N}(\boldsymbol{\mu}_x, \Sigma_x)$, and $\mathbf{P_y} = \mathcal{N}(\boldsymbol{\mu}_y, \Sigma_y)$. Σ_x and Σ_y are diagonal matrices, i.e. $\Sigma_x = diag(\boldsymbol{\sigma}_x)$, and $\Sigma_y = diag(\boldsymbol{\sigma}_y)$. $\mathbf{0}_p$ and $\mathbf{1}_p$ are p-dimentional 0 and 1 vectors. For every parameter setting, the following experiment is repeated 200 times. $\mathbf{X}$ and $\mathbf{Y}$ are drawn and MMD and classification accuracy are computed. The empirical null distribution for MMD and classification accuracy is obtained with a permutation test of 500 permutations, and used to reject the null with $\alpha = 0.05$. The median heuristic is used with the MMD test, and the width of the SVM kernel is selected by cross-validation. [Top Row] $\mathbf{P_x}$ and $\mathbf{P_y}$ vary in their means. [Bottom Row] $\mathbf{P_x}$ and $\mathbf{P_y}$ varry in their standard deviation. [Left] Power for different values of the sample size n. [Middle] Power for different values of p ($\mathbf{P_x}$ and $\mathbf{P_y}$ vary along only 5 dimensions to avoid trivial rejection). [Right] Power for different parameters of the distribution.

These results are far from being a proof, as one might be able to construct a more complex

classifier that would beat MMD in power. However, this is precisely our point. Because the truth is not known, it is difficult to optimize the power of a test, as there is a very large search space to go over with no clear indication of how to proceed. A much simpler alternative is to use the MMD, which has theoretical foundation and has been proven to be consistent for two sample testing [43].

6.2 Independence testing



When experimental stimuli do not fit neatly into conditions and are of a complex nature (e.g. pictures or words), it is not useful to perform a two-sample test, and people use encoding (or predictive) models [81, 86]. While decoding models express the label of the stimulus as a function of the underlying activity, encoding models express the brain activity as a function of the stimulus. This is done through an intermediate feature space (IFS) of the stimulus (e.g., if the stimulus is an image, an IFS can be a vector representing some visual properties of the image). The next step is to predict the brain activity as a function of the different elements of the IFS just as we showed in chapter 3. As we discussed in chapter 2, the IFS approach is very useful for studying brain representations: using the intermediate feature representation of stimuli, a predictive model is able to generalize from a finite set of examples and predict the activity for unseen stimuli. Whereas the decoder from the previous section can only distinguish between classes seen in training, an IFS approach allows the experimenter to abstract the stimuli into a more general and much more expressive space.

There are multiple ways to judge the encoding model's predictive performance. Encoding models can be used for decoding like we did in chapters 3, 4 and 5: a classifier is asked to guess which of the stimuli a and b corresponds to a given brain image $\mathbf{v}$. Using the IFS, it predicts the activity for the two stimuli: $\mathbf{v}_a$ and $\mathbf{v}_b$. It assigns to $\mathbf{v}$ the label a or b depending on which of $\mathbf{v}_a$ and $\mathbf{v}_b$ is closest. Average classification accuracy is tested for significance with a permutation test. There are other statistics for testing the encoding model, such as computing the percentage of variance explained on held-out data [86].

Let's rephrase the main operating strategy behind IFS analysis, which we have been using throughout this thesis: IFS analysis can be used to identify where in the brain the properties of the stimuli are represented. One can use multiple IFSs to represent the same stimuli (for example, use an IFS that represents the visual properties of images and another IFS that represents their

semantic content). An encoding model can then be used to find regions in the brain that are related to each IFS. The working assumption is that, if a region in the brain is not representing a given type of feature (for example semantic properties), then the encoding model will fail at predicting brain activity correctly and the classification accuracy will not be significantly higher than chance. On the other hand, if the classification accuracy is higher than chance in a region r using an IFS j (e.g. visual IFS), then this suggests that region r is processing the corresponding type of properties (e.g. visual information).

A major observation of this chapter is that IFS analysis is an indirect **independence test**. Given a set of paired observations $\mathbf{x}_i$ and $\mathbf{y}_i$, an independence test tries to find if $\mathbf{P}_{XY} \neq \mathbf{P}_X \times \mathbf{P}_Y$. In our case, given a set of paired observations of some properties of a stimulus $\mathbf{f}_i^j$ and brain activity $\mathbf{v}_i^r$, we would like to know if region r is representing the properties of type j. In order to evaluate that, we test if the activity in region r and the properties of type j are related, i.e. we want to know if $\mathbf{P}_{F^j V^r} \neq \mathbf{P}_{F^j} \times \mathbf{P}_{V^r}$.

The encoding approach we described previously in this thesis is quite complex. It necessitate first the estimation of regression function, and then a classification task. Regression is known to suffer from the curse of dimensionality and is not reliable for high dimensional variables in a small sample setting. As in Section 6.1, we might suffer an estimation bottleneck while we only need a test. Additionally, the classification task necessitates multiple decisions, such as the ones made for the cross-validation setup.

We propose to perform the independence test **directly**. We need a non-parametric test that does not require prior assumption of the form to the joint and marginal distributions $\mathbf{P}_{XY}$, $\mathbf{P}_X$ and $\mathbf{P}_Y$. We need this method to be multivariate, to test for independence of multi-dimensional variables that differ in their size. The Hilbert-Schmidt Independence Criterion (HSIC) [38], is a kernel independence test statistic that has been developed for these purposes. The population HSIC is equal to 0 if and only if X and Y are independent (for details refer to [38]). The HSIC tests:

$$H_0 : \mathbf{P}_{XY} = \mathbf{P}_X \mathbf{P}_Y \quad \text{against} \quad H_1 : \mathbf{P}_{XY} \neq \mathbf{P}_X \mathbf{P}_Y.$$

Given characteristic kernels k and l, the matrices $\mathbf{K}$ and $\mathbf{L}$ are the $n \times n$ sample kernel matrices defined such that $\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathbf{L}_{ij} = k(\mathbf{y}_i, \mathbf{y}_j)$. $\mathbf{H} = \mathbf{I_n} - \frac{\mathbf{1}}{\mathbf{n}} \mathbf{1_n} \mathbf{1_n}^\top$ is a normalizing matrix and $\tilde{\mathbf{K}} = \mathbf{HKH}$ and $\tilde{\mathbf{L}} = \mathbf{HLH}$ are normalized kernel matrices. The sample HSIC is computed as

$$\text{HSIC}_n = \frac{1}{n^2} \text{tr}(\tilde{\mathbf{K}} \tilde{\mathbf{L}}) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \tilde{\mathbf{K}}_{i,j} \tilde{\mathbf{L}}_{i,j},$$

In practice, we reject the null hypothesis if HSIC_n is "close" to 0. We estimate the threshold with a permutation task. This permutation task simulates the null hypothesis by permuting the samples of $\mathbf{X}$ into $\mathbf{X}'$, making the samples $\mathbf{x}_i'$ and $\mathbf{y}_i$ unrelated. This is done a large number of times and HSIC_n is computed. These samples are then used to reject with level α the unpermuted HSIC_n statistic.

Notice how simple the setup of the HSIC test is in comparison with training an encoding model and then performing classification. After constructing and normalizing the kernel matrices, a simple element-wise product is performed, followed by computing the average of all the resulting matrix elements. The permuted statistics are also easy to compute since one doesn't

need to estimate another kernel matrix, but only has to permute $\tilde{\mathbf{K}}$ appropriately. In our experiments, we choose k and l to be Gaussian kernels, and we pick the parameter $2\sigma^2$ for each kernel using the median heuristic [113], i.e. $2\sigma^2$ is set to the median of the square distances $||\mathbf{x}_i - \mathbf{x}_j||^2$ and $||\mathbf{y}_i - \mathbf{y}_j||^2$ respectively, for $i \neq j$. While linear correlation only checks for linear dependence, the Gaussian-HSIC is *consistent* (for any choice of bandwidth), i.e. given enough paired samples $\mathbf{x}_i$ and $\mathbf{y}_i$, this test will detect any differences in $\mathbf{P}_{XY}$ and $\mathbf{P}_X \times \mathbf{P}_Y$.

Another approach to studying fMRI representations, called Representation Similarity Analysis (RSA) [66], has quickly gained popularity [64]. RSA is the closest fMRI method to a clear independence test for brain and feature representations (although RSA is not explicitly referred to as a statistical independence test). The RSA relies on the assumption that brain regions and features that encode the same type of information should have a similar structure of inter-stimuli distances. RSA uses dissimilarity matrices, which correspond to 1 minus the correlation matrix of the sample set[2]. Dissimilarity matrices are computed using different IFSs, as well as using the activity in different brain regions. A dissimilarity matrix encodes the similarity of object representations in different spaces. Dissimilarity matrices for IFS j and brain region r are compared[3]. Brain regions and features that encode the same type of information should have a similar structure of inter-stimuli distances. The significance of this measure is assessed via a with a permutation test.

The computations behind RSA resembles enormously the HSIC test. However, unlike HSIC, RSA doesn't have theoretical guarantees of consistency and can reveal only linear similarities. Nonetheless, RSA gives us an intuitive explanation to why HSIC is a measure of dependence – the matrix $\tilde{\mathbf{K}}$ measures how similar the $\mathbf{x}_i$s are to each other (same for $\tilde{\mathbf{L}}$ and $\mathbf{y}_i$s); if $\tilde{\mathbf{K}}$ and $\tilde{\mathbf{L}}$ have similar structure (large matrix dot-product, as captured by HSIC$_n$), then X and Y must be "similarly similar" and hence dependent.

6.3 Accounting for naturalistic experiments with continuous recording

The HSIC permutation test simulates the null distribution under the assumption that the pairs $(\mathbf{x}_1, \mathbf{y}_1)...(\mathbf{x}_n, \mathbf{y}_n)$ are independent and identically distributed (IID). However, when a naturalistic experiment consists in stimuli that are not neatly divided into trials and the recording is continuous (e.g. when the subjects watch a movie or read a story such as in this thesis), the resulting data is a time series, and the samples obtained are not IID. If $\mathbf{x}_1...\mathbf{x}_n$ and $\mathbf{y}_1...\mathbf{y}_n$ are times series, running a permutation test under-estimates the breadth of the null distribution, which can result in a high rate of false positives[4]. In [14], a new test is proposed that empirically estimates the null distribution by circularly shifting $\mathbf{X}$ by c time points for $a \leq c \leq b$ before computing HSIC, where a is chosen large enough so that the dependence between Y_{t+a} and X_t is negligible,

[2]This quantity is dubbed "correlation distance", not be confused with "distance correlation" [118].

[3]Pairs of matrices are compared by looking at the Spearman rank correlation of their upper halves.

[4]To see this, consider that when both variables X_t and Y_t are random processes with successive samples that are correlated, i.e. both their sample kernel matrices will have a clear structure around the diagonal because nearby points are similar. Therefore, the element wise product will be high, and HSIC will be higher for the unpermuted sample than for the permuted sample, even when there is no real dependency between X_t and Y_t.

and b is smaller than the sample size. This shifting procedure is shown to correctly estimates the null distribution by maintaining the temporal nature of the samples while at the same time destroying the relationship between X_t and Y_{t+a} due to the large time lag.

Shift-HSIC [14] therefore enables independence testing of the fMRI activity V_t^r in a brain region and the feature representation F_t^j. This allows the following interesting manipulation. FMRI measures the change in the oxygen level of the blood that occurs as a result of neural activity, i.e. the hemodynamic response. We saw in chapter 3 that this response has a latency of several seconds (see Fig. 6.2 [Left]). Because of this latency, one should not only look for dependencies between F_t^j and F_t^j, but also for dependencies between F_t^j and V_{t+k}^r, where $k \in \{0, ..., k^*\}$ and k^* is close to the length of the hemodynamic response. This procedure is summarized in Fig. 6.2 [Right].

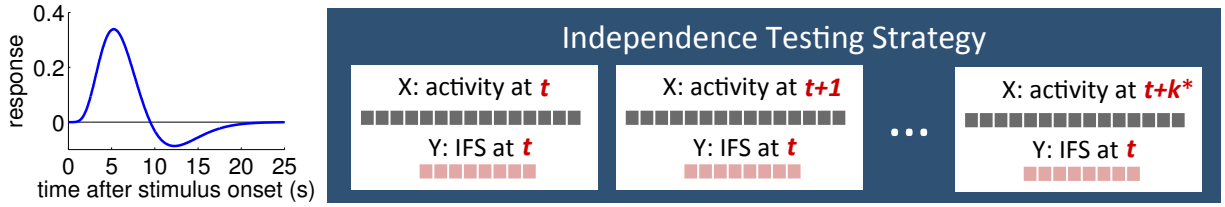


Figure 6.2: [Left] A typical hypothetical hemodynamic response. The brain activity response is delayed with respect to stimulus presentation. [Right] Independence testing strategy. The test is repeated at various delays of the brain activity to detect lags in dependence due to the slow dynamics of brain activity.

Our procedure (1) takes into account the inherent lag in the fMRI signal and (2) might lead to interesting findings if different feature types affect the fMRI activity with different latencies. In our experiments, we pick $k^* = 6$, which corresponds to a lag of 12 second. **Remark**: if we are not interested in figuring out the exact latency, we can alternatively combine the $\mathbf{f}_t^j$ samples and create $\mathbf{f}_t^{j'} = [\mathbf{f}_{t-T}^j, ..., \mathbf{f}_{t-1}^j, \mathbf{f}_t^j]$, and test the independence of $F_t^{j'}$ and V_t^r, similar to the analysis we did in chapter 4 and in [127].

6.4 Real data experiment

We replicate here the experiment in chapter 4. Remember, 8 subjects read a chapter of *Harry Potter and the Sorcerer's Stone* while words were presented one at a time. The story consisted in 5176 words presented in 4 experimental blocks. The total presentation time is 45 minutes and the available data consists of 1355 volumes of fMRI activity for each of the 8 subjects, each scanned with the same timeline of stimulus presentation. The subjects brain contain 30,000-40,000 voxels, of size $3 \times 3 \times 3$mm^3, which are all normalized to the same coordinate space. The intersection of the locations of all the voxels across subjects consists of 41,073 unique locations. We used for this experiment the non spatially-smoothed version of the data[5], that we then preprocessed with all the later steps used in chapter 4[6]. Additionally, we threw out the first and last 20 TRs from

[5]Remember the data is available at `http://www.cs.cmu.edu/~fmri/plosone`
[6]All the results in chapter 4 are also provided for this non-smoothed dataset in Appendix C.

each blocks so that the starts and ends of blocks don't introduce spurious relationships between the IFS and the data.

We use here the semantic IFS annotations from chapter 4. We also use the concatenated searchlight procedure from chapters 3 and 4: the searchlight procedure from [65] is modified to include data from all participants. At every iteration, a cube of voxels is selected (we use $5 \times 5 \times 5$ voxel cubes, i.e. cubes of $1.5 \times 1.5 \times 1.5$cm^3). This cube is moved over the entire brain, so that it's centered at every possible voxel location. For each cube location, we include the voxels from all the subjects that belong to this cube. As argued previously, this procedure pools data together from all the subjects without requiring their brain to align perfectly to each other.

Classification Accuracy Statistic:

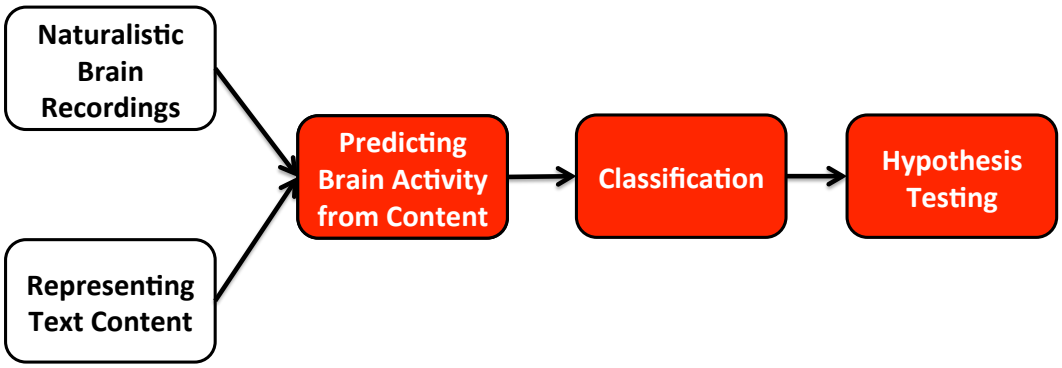


Figure 6.3: Previous IFS analysis pipeline.

We replicate here the investigation pipeline from chapters 3 and 4. We will summarize here these steps again for ease of comparison with the other method we propose. Mainly, this analysis consists in a 10-fold cross-validation of the following steps. For every voxel location r, we consider the vector of brain activity $\mathbf{v}^r$ formed of the voxels in cube r:

a. Using 90% of the data, fit a predictive model of $\mathbf{v}^r_{t+k}$ as a function of the semantic IFS $\mathbf{f}_t$. We use ridge regression with a penalty chosen by nested cross-validation for each voxel independently.

b. Divide the test samples into non-overlapping 40s segments. For each 40s segment:

   i. Predict the activity for the corresponding 40s feature segment, and an incorrect 40s segment.

   ii. Compute the Euclidean distance between the two predicted segments and the real segment over the voxels included in the cube r. Pick the label of the closest passage.

c. Repeat for all folds and compute average classification accuracy.

In the original approach, we accounted for the delay in the hemodynamic response by delaying and combining the feature vectors into $\mathbf{f}_t{}' = [\mathbf{f}_{t-4}, \mathbf{f}_{t-3}, \mathbf{f}_{t-2}, \mathbf{f}_{t-1}]$. Here, we test the dependencies between fMRI data at time $t+k$ and the text features at time t, independently for each $k = \{0...6\}$.

HSIC Statistic:

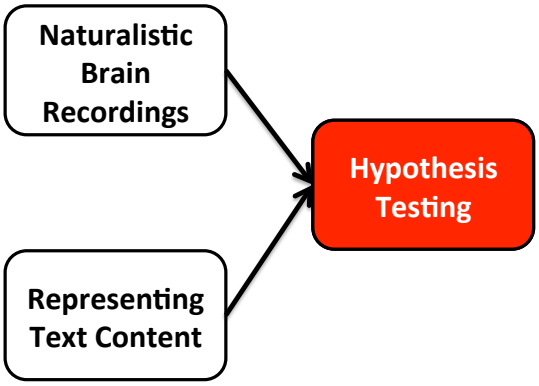


Figure 6.4: Proposed simpler pipeline.

We contrast this complex classification procedure with an HSIC test. For each delay k, we compute the HSIC of the samples $\mathbf{f}_t$ and the brain activity $\mathbf{v}^r_{t+k}$ of searchlight cube r.

Shift test

To assess a p-value for one of our statistics U (HSIC or classification accuracy), we compute an empirical chance distribution for the null distribution using a shift test. The null hypothesis is that the brain activity in a searchlight cube r and the semantic features are independent.

- We delay $\mathbf{v}^r$ by $\{150, 151, 152...1149\}$ TRs, totaling 1000 delays. 150 is a large enough lag that the brain activity and the semantic features should be unrelated. We compute the statistics $\{U^{\text{null}}_{r,d'},\ 150 \leq d' \leq 1149\}$ that approximate the null distribution.
- For every location r, we have 7 statistics of interest: $\{U_{r,d},\ 0 \leq d \leq 6\}$, corresponding to times $\{0s, 2s, ...12s\}$ after stimulus onset. We compute the p-value $p_{r,d}$, i.e. the proportion of $\{U^{\text{null}}_{r,d'},\ 150 \leq d' \leq 1149\}$ that is at least as extreme as $U_{r,d}$, for $0 \leq d \leq 6$.
- We use the Benjamini-Hochberg false discovery rate procedure [6] to control for multiple comparison involving all r and d. It is important to state that for both the HSIC and the classification accuracy test, **the false discovery rate is controlled at q** (we use $q = 0.05$).

Results

Fig. 6.5 shows the results for the two tests, and is meant as a demonstration of the similarity of the two procedures, although the HSIC test is considerably simpler. The pattern revealed by the HSIC test appear to be smoother, but it's difficult to ascertain which pattern is closer to the truth. However notice how the voxels that are detected by the HSIC method are mostly "on" for 4 to 8 seconds, i.e. when we expect the hemodynamic response to the semantic features presentation to take place. In comparison, the voxels identified by the classification accuracy method are not as consistent in time.

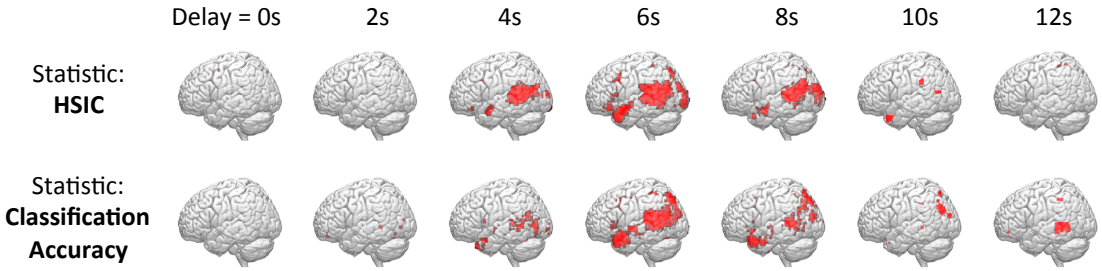


Figure 6.5: Dependencies between the fMRI signal of voxel cubes centered at different locations and the semantic features of words in the stimulus text, at various delays after the features are presented, as revealed using the HSIC or the classification accuracy tests. The locations of the centers of the cubes with dependent signals are show in red after controlling the false discovery rate at $q = 0.05$.

6.5 Discussion

Our framework encompasses the analysis of controlled and naturalistic experiments under the umbrellas of two-sample and independence testing. We advocate the use of direct, non-parametric kernel hypothesis tests that adapt readily to multivariate fMRI data and IFS representations. Our experimental results suggest that the power of direct kernel hypothesis tests is as least as good as performing an indirect test where a regression or classification model is trained on the data and then used to generate a test statistic.

The variety and flexibility of regression and classification techniques are, in our view, more of a bane than a boon for such high-dimensional, low-sample size experiments with high degrees of freedom, causing experiment results to vary based on many details like which regression or classification technique was used, what regularization penalty or model selection procedure was used (like cross validation, AIC, BIC, etc), etc. In addition, they are also solving a harder problem (estimation) as an intermediate step to the desired goal (testing), and often form test statistics after estimation based on classification accuracy, prediction errors, etc. In contrast, our direct tests require (almost) no parameter choices, don't rely on the choice of any generative model, or on assumptions like a linear model or Gaussianity on the signal/noise, etc.

By advancing a clear framework for neuroimage analysis problems and stating them as hypothesis tests, we believe this chapter also contributes in making naturalistic brain image analysis more accessible to both neuroscientists and machine learners. The reading experiment in chapter 4 is a complex experiment consisting of complex stimulus that varies along an infinity of dimensions (semantic properties, syntax, narrative structure, etc.). It is also a single trial experiment, which is a very rare experimental design (it is preferred to have many repetitions of the stimulus to increase signal to noise ratio by averaging). Nevertheless, we were able to clearly state the original complex problem presented in 4 and [127] as an independence test between different voxel locations and the IFS of the text. By incorporating the effect of the hemodynamic response as a delayed dependence between the brain activity and the stimulus features, we were able to easily incorporate in our framework naturalistic experiments with continuous fMRI recordings that are not neatly divided into trials.

Future Work: The tests we advocate for are consistent for any choice of bandwidth of the Gaussian kernel, or any other characteristic kernel, although different choices will affect the test power. The few choices left to be made are therefore important, but since the framework is simplified, most of the experimentalist's effort can be concentrated on finding appropriate kernels for brain images and stimulus features, as well as kernel parameters that maximize the power of the associated test. Meanwhile, the Gaussian kernel with the median heuristic (see Section 6.2) serves as a natural and popular default choice. As an example, one interesting way to *learn* a good kernel is to use a subset of the experiment for "calibration": in that subset, stimuli are repeated and the kernel is optimized to give high similarity to brain activity from presentations of the same stimulus.

This chapter lays the ground for crucial future work. In naturalistic experiments, the different IFSs of the stimuli might be correlated. The marginal dependency between a region and an IFS might therefore be due to spurious correlations with another IFS. As we saw in chapter 5 (were we tried to offer an estimation-based test for *conditional* independence) functional neuroimaging will benefit enormously from non-parametric and multivariate *conditional* independence tests. Such direct tests exist [31] for limited settings, and problems exist in using them for non-IID, high-dimensional, real-valued variables, especially (but not only) when it comes to performing a correct permutation test. Therefore, progress on statistically correct conditional independence tests for non-IID processes will be of utmost importance for functional neuroimaging.

In Appendix E, we include the results from a paper in which we found that using Stein Shrinkage when estimating HSIC may improve the power of the independence test. We propose two shrunk HSIC estimators. These estimators might be useful in analyzing functional neuroimaging recordings.

Chapter 7

Discussion

This thesis presented an integrated approach to study reading and language processing in the brain using naturalistic experiments, and offered a hypothesis of what processes are computed in different regions of the brain, by modeling the content of the text using semantic vector space models, syntactic information, hand labelled narrative structure annotation and vectors derived from neural network models. This thesis was a methodological effort accompanied by novel results, and we revisit here the disciplines to which it has contributed.

7.1 Cognitive neuroscience of language

We have presented and advocated for a naturalistic approach to study language processing. We used a single fMRI study to replicate multiple results from the language processing literature (where each result typically required a separate experiment). These results provide a solid proof-of-concept for our claim of studying brain activity using a complex natural task. Furthermore, we used Magnetoencephalography (MEG) in order to uncover the rapid dynamic processes involved in perceiving the properties of consecutive words and representing their continuously updated context. Our novel results suggested a spatiotemporal map of the progressive perception of a word in the brain, which can be consolidated and refined in future work.

7.1.1 Cognitive neuroscience methodology contributions: a naturalistic imaging task with no repetitions

We have shown that it is possible to analyze data from a naturalistic experiment in which the text was not controlled, using a slow imaging technique like fMRI, and a noisy, low spatial resolution technique like MEG. We were able to build a model of the complex task of processing of natural text, as a function of the properties of the text.

Our methods did not require having repetitions of stimuli. Most fMRI and MEG experiments repeat stimuli and average the repetitions to improve signal to noise ratio. Since the length of experiments is usually limited, this reduces significantly the variability of the stimulus that can be presented. Having no repetitions allows more varied stimulus, and we have shown that we can overcome the lack of repetitions as long as the language features of interest occur with some

frequency in the text, and with enough variability across features that their contributions to the signal can be disentangled.

We believe that freely sharing data online is a great way to encourage replication of results. It is also a great way of building our common knowledge of the brain faster by "crowdsourcing" the efforts of multiple investigators worldwide, as the problem we are solving is very complex and has lots of room for investigation. We have made the data from the fMRI experiment available online[1], and will make the MEG data available soon. The language features we use are also made available on the same website. Any researcher is thus welcome to download the data to test our language features, test their own language features, or contrast different theories of language.

While this thesis studies language processing, our approach does not have any core requirements that prevents it from being used to study other cognitive tasks. The central idea is to represent the processes at hand in an intermediate feature space (IFS). In [91], the stimulus is natural videos and the IFS is a set of motion features from this video. In [61], the IFS for these videos consists in semantic annotations of the objects that appear in the different scenes. Therefore, we know that approaches similar to ours can be used when studying vision. We suggest that this approach can be used to study many high-level cognitive tasks. The idea is that for any cognitive task of interest that you perform in a naturalistic manner (e.g. maintaining a conversation, writing, listening to music, various cognitive tests) once it is determined that you can establish an IFS of the tasks being performed, finding their relationship with the brain activity of different regions seems like an achievable next step. Think of the majority of current fMRI experiments: to study a cognitive task α (e.g. music perception), an experiment consists of a small number of conditions (e.g. pleasing consonant music vs. displeasing dissonant music, see [63]), and the effect of each condition on the brain is measured to see where in the brain a condition's effect is different from baseline, or two conditions have different effects. As we saw in chapter 6, this is a *two-sample* problem. With our approach, we are proposing to do a naturalistic experiment that would constitute the *independence test* that is analogous to that two-sample test, in order to study the same task α. The hope is that this approach will be able to uncover rich brain representations as we saw in the reading case, and offer new directions for close investigation with additional controlled experiments.

For some cognitive tasks, the difficulty might be in building such an IFS. It might be difficult when studying tasks like retrieval from memory, complex decision making or problem solving to pinpoint the actual timings of the different underlying mental processes. However, this has not stopped researchers from attempting to find the signatures of these processes in an unsupervised manner. In [7] the authors uncover, from the properties of EEG activity, three different stages that are related to an associative recollection task, during which the subjects were asked to recognize if they had seen a word pair in a previous study phase. These identified stages have variable durations and are characterized by learned distributions. In [60], the onset of different processes involved in a question answering task was also estimated from data. In general, an important line of investigation is to learn from the data the different relevant IFSs in an unsupervised way. This would help in the case of tasks that are hard to model with an IFS because the exact processes of interest are unknown, but also in cases where multiple IFSs can be built such as our reading problem. In such cases, we do not know the appropriate IFSs that we should use, and uncovering

[1] `http://www.cs.cmu.edu/~fmri/plosone/`

them automatically would offer a less biased illustration of the underlying mental processes.

7.1.2 Results: the start of a spatio-temporal reading map

The fMRI experiment has allowed us to build brain representation maps that distinguish different brain areas based on what type of activity they are processing. The MEG experiment completed the picture by drawing a rich spatio-temporal progression of the representation of different features in the brain. It also revealed how the processing of a word lasts after the next word is on the screen, and where and when the representation of the previous context is stored in the brain. We are not including in this section the actual list of regions we have identified and their precise location, because we see the work we have presented in this thesis as a stepping stone to a long series of investigations and refinements. The use of more expressive NLP tools that provide a more appropriate IFS, denoting for example sentence meaning (instead of single word meaning), will lead to more complete results, and so will future models of single word semantics, syntax etc. The analytical tools of the future might also be more powerful for dealing with noise, for detecting real dependencies from spurious correlations or for modeling subject specific variations in a more comprehensive and useful manner. For these reasons the representation maps we have provided are far from complete. Moreover, we see the construction of these map as necessitating a back and forth between controlled experiments and naturalistic experiments to reinforce and test its different parts.

We have seen in chapter 2 multiple models of sentence processing, such as Friederici's cortical language circuit. In [27], Friederici proposes a timeline for each of the steps involved in auditory sentence comprehension, including the occurrence of semantic processing and syntactic processing in parallel between 300-500ms after word onset, and syntax and semantic integration around 600ms. Our results suggest however that semantic features activate well before 300ms. There might be many reasons for this disparity, especially since we are comparing different tasks (reading and auditory comprehension). However, rather than a difference in the modality that language is perceived in, the fact that we find semantic features to be represented earlier *might* be due to the core of our philosophical approach. Instead of "breaking" a language process, such as presenting a semantically incongruous word or syntactically unexpected word (which is the investigation strategy used to construct the model in [27] and most other models of language), we instead study the brain as it perceives language normally, by modeling the content of what it is perceiving. The perception of a word has to occur before the process that determines if this word is congruous or not. By using a semantic incongruity task to judge semantic perception, one might therefore suffer from a timing bias. In general, when using an approach that breaks normal language processing in order to study normal language processing, we might introduce timing biases as well as others biases. Naturalistic experiments escape this risk, and should be seriously considered as a powerful investigation method.

7.2 Statistical Machine Learning

To study natural reading, this thesis relies on finding relationships between the brain activity at different locations and timings and the different properties of the rich stimulus text being read.

We have seen that this is a challenging statistical problem. On one hand, brain data has a very large number of dimensions, has rich spatiotemporal structure, has a low signal to noise ratio and is typically acquired in 1-2 hour experimental sessions, leading to a small number of samples when compared to the large dimensionality. On the other hand, natural text varies along a very large number of dimensions, and these dimensions are highly likely to be correlated (especially in a small sample setting), which creates a risk for the relationships we find between brain activity and a text dimension being confounded with other dimensions. In this thesis we have tried to build some appropriate computational methods. We have presented a complex investigation pipeline that we used to provide unprecedentedly rich spatiotemporal brain maps of language processes. We have continuously and closely inspected our methods to enforce robustness of results, and we consider that this investigation pipeline will always be a subject of continuous research.

For instance, towards the end of this thesis, we have even proposed an alternate, simpler pipeline to our entire computational approach, as well as important directions for future work to extend it. In that chapter (chapter 6), we looked at commonly used techniques like brain decoding, encoding models and representation similarity analysis from the perspective of hypothesis testing. We introduced different tests to be used and have shown experimentally that they seem to work at least as well as the more complicated current alternatives. Many of the current alternatives require a model assumption and estimation, while the existing kernel hypothesis tests we suggest to use are non-parametric, i.e. they will not suffer from an incorrect model assumption or estimation.

No-repetition imaging: As we mentioned previously, we have shown in this thesis that it is possible to use a no-repetition paradigm, which the IFS approach makes possible, as well as various procedures we used to improve the signal to noise ratio. For instance in the classification task in sections 3, 4 and 5, we improve the low signal to noise ratio by grouping words together in testing, and by concatenating the brain images of subjects that are processing the stimulus at the same time. We showed in chapter 5 that this improves accuracy significantly. In chapter 6, we also use the concatenation of data from multiple subjects in the aim of increasing the power of our tests.

The clear expression of our approach as an independence test in chapter 6 allows the problem to be more directly and quickly accessible for researchers from various disciplines, and will hopefully motivate the development of new computational methods that are statistically sound. However, one should always exercise caution when dealing with a statistically challenging problem such as ours: it is not always clear how to correctly formulate the null hypothesis or how to deal with correlations. Moreover, one should always be careful when engaging in statistical testing, since if one keeps testing various hypotheses on the same dataset, one hypothesis is bound to appear true by chance. Replicating the findings with new stimuli and new data is required to add confidence to results.

7.3 Natural Language Processing

This thesis does not present new work in NLP, however, it offers a new way of testing and comparing different NLP algorithms and language models. We have shown how we use various

NLP models to find a numerical representation of the semantic features of individual words, their syntactic properties, the context of the previous words etc., and then use these numerical representations as an IFS to study brain activity. We chose these specific models because we considered them appropriate, but the approach we presented in this thesis is agnostic to the model that is chosen. Because NLP models are not yet able to capture the entire complexity of language, a specific model we use in our approach will always be an interim model that can be changed for another, better model that proves to be more accurate at predicting brain activity.

Our approach can be applied to any model of language as long as features can be extracted from that model. This means, after the experimental data is collected with a rich text, the experimenter can go back and analyze it again with a new model of language processing, without having to perform another experiment (again, care should be taken not to fall into cherry picking fallacies and confirmation biases; this could be achieved for instance by partitioning the dataset into a training set and a untouched test set on which models are finally compared[2]). In the future, the brain recordings of subjects reading natural text could therefore present a new method for deciding which of several models is more appropriate (as measured by its similarity to brain computations). These recordings might even be an additional source of data for the training of NLP algorithms, so that they arrive at more cognitively plausible language representations.

7.4 Future Work

We present in this section different ways this thesis can be extended:

Conditional independence testing: We are working on replacing the classification test that we presented in 5 with a conditional independence test that can be shown to be consistent.

Source localization of the MEG dataset: We are currently working on source localizing the MEG dataset in order to investigate more closely the location of the different reading processes we obtained. This would allow us to better compare the results with the results of the fMRI experiment.

Combining fMRI and MEG: The next logical step in our investigation seems to be using fMRI and MEG together to benefit from both the high spatial resolution and the high temporal resolution they offer respectively. We are currently working on a model that estimates the latent space of neural activity from both datasets jointly. We take the fMRI and MEG recordings of the same person reading the same text with the same presentation timings. Both of these datasets are distorted and smoothed versions of the underlying activity (which has high resolution in space and time). We express fMRI and MEG recordings as functions of the underlying activity and we jointly estimate these functions and the underlying neural activity. The hope is that this approach will allow us to better localize brain processes in space and time.

Joint learning of statistical language models and brain activity predictors: Another extension we are currently working on is to build neural network models of language that, after an initial training on a corpus, are trained to predict both the MEG activity and the incoming word.

[2]Since most of the time the data would have already been used in its entirety, the validation of findings about the brain is going to eventually require a new dataset. However, if multiple experimenters use this kind of naturalistic paradigms and share their data, more and more data will be available for such purposes. Data from such experiments will not be constrained by a specific task and it will be possible to use it to test various models about language.

There are various interesting setups of models that we can try, which may or may not include the brain activity from the previous word as input to the model. A very interesting outcome of such approaches is to eventually be able to build statistical language models that perform better at NLP tasks (and not just brain prediction tasks) after the inclusion of brain data in training. This however seems like a hard problem because of the limited sample size of brain imaging data, especially when compared to the size of the corpora that statistical language models typically need in training. However, this research direction has a very nice intellectual appeal: the statistical language model is trying to perform a task that the brain is able to perform, and therefore a better knowledge of how the brain works *might* lead to better language models[3].

Studying individual differences: It would be very interesting to study the variability of the spatio-temporal maps across different individuals. An extension of our work that *might* have useful practical application is to develop methods to characterize what is common to the brain maps of readers in the same population (for example good readers, or readers with a specific type of dyslexia), and what varies between populations. This would be a challenging statistical problem and would require rigorous statistical methods to distinguish individual variations from noise or spurious correlations. However, the potential of this approach might be very useful for better understanding reading disabilities: representation maps might help us understand individual differences in behavior. For example, we might be able to diagnose the different types of dyslexia from a brain scan. We might understand better the differences between the brains of good readers and the brain of a student with a reading problem, and propose individually tailored educational strategies.

[3]As we saw in chapter 2, [32] does a first, encouraging step in this direction.

Appendices

Appendix A

Methods for learning brain responses

A.1 Small Area model and Gibbs sampler

Here, we specify the details of the Gibbs sampler for the hierarchical Bayesian small-area model described in section 3.5.2. Recall that $A(v)$ is the area which voxel v resides in, and the $V(a)$ the set of all voxels in area a.

The joint distribution can be written as

$$
\begin{aligned}
P(\mathbf{Y}, \mathbf{U}, \beta, \alpha^2, \nu^2, \sigma^2) = \prod_a \mathcal{IG}(\alpha_a^2|c,d) \times \prod_v \mathcal{IG}(\nu_v^2|e,f) \times \prod_v \mathcal{IG}(\sigma_v^2|a,b) \\
\times \prod_a \mathcal{N}(\mathbf{u_a}|0, \alpha_a^2\mathbf{I}) \times \prod_v \mathcal{N}(\mathbf{z}_v|0, \nu_v^2\mathbf{I}) \times \prod_v \prod_t \mathcal{N}(y_{vt}|(\mathbf{u}_{A(v)} + \mathbf{z}_v)^\top \mathbf{x}_t, \sigma_v^2).
\end{aligned}
$$

We now derive the full conditionals from the joint distribution above.

$$
\begin{aligned}
P(\mathbf{z}_v|...) &\propto \mathcal{N}(\mathbf{z}_v|0, \nu_v^2\mathbf{I}) \times \prod_t \mathcal{N}(y_{vt}|(\mathbf{u}_{A(v)} + \mathbf{z}_v)^\top \mathbf{x_t}, \sigma_v^2) \\
&= \mathcal{N}(\mathbf{z}_v|\boldsymbol{\mu}_{\mathbf{z}_v}, \boldsymbol{\Sigma}_{\mathbf{z}_v}), \\
\boldsymbol{\Sigma}_{\mathbf{z}_v} &= \left(\frac{1}{\nu_v^2}\mathbf{I} + \frac{1}{\sigma_v^2}\mathbf{X}^\top\mathbf{X}\right)^{-1} \\
\boldsymbol{\mu}_{\mathbf{z}_v} &= \boldsymbol{\Sigma}_{\mathbf{z}_v} \times \frac{(\mathbf{X}^\top\mathbf{y_v} - \mathbf{X}^\top\mathbf{X}\mathbf{u}_{A(v)})}{\sigma_v^2}.
\end{aligned}
$$

$$
\begin{aligned}
P(\mathbf{u}_a|...) &\propto \mathcal{N}(\mathbf{u_a}|0, \alpha_a^2\mathbf{I}) \times \prod_{V(a)}\prod_t \mathcal{N}(y_{vt}|(\mathbf{u}_{A(v)} + \mathbf{z}_v)^\top \mathbf{x_t}, \sigma_v^2) \\
&= \mathcal{N}(\mathbf{u}_a|\boldsymbol{\mu}_{\mathbf{u}_a}, \boldsymbol{\Sigma}_{\mathbf{u}_a}), \\
\boldsymbol{\Sigma}_{\mathbf{u}_a} &= \Big(\frac{1}{\alpha_a^2}\mathbf{I} + \mathbf{X}^\top\mathbf{X}\sum_{V(a)}\frac{1}{\sigma_v^2}\Big)^{-1} \\
\boldsymbol{\mu}_{\mathbf{u}_a} &= \boldsymbol{\Sigma}_{\mathbf{u}_a}\sum_{V(a)}\frac{(\mathbf{X}^\top y_v - \mathbf{X}^\top\mathbf{X}\mathbf{z}_v)}{\sigma_v^2}\ .
\end{aligned}
$$

$$
\begin{aligned}
P(\sigma_v^2|...) &\propto \mathcal{IG}(\sigma_v^2|a, b) \times \prod_t \mathcal{N}(y_{vt}|\boldsymbol{\beta}_v^\top \mathbf{x_t}, \sigma_v^2) \\
&= \mathcal{IG}(\sigma_v^2|a', b'), \\
a' &= \frac{2a + T}{2} \\
b' &= \frac{2b + (\mathbf{y}_v - \mathbf{X}\boldsymbol{\beta}_v)^\top(\mathbf{y}_v - \mathbf{X}\boldsymbol{\beta}_v)}{2}.
\end{aligned}
$$

$$
\begin{aligned}
P(\alpha_a^2|...) &\propto \mathcal{IG}(\alpha_a^2|c, d) \times \mathcal{N}(\mathbf{u_a}|0, \alpha_a^2\mathbf{I}), \\
&= \mathcal{IG}(\alpha_a^2|c', d') \\
c' &= \frac{2c + P}{2} \\
d' &= \frac{2d + \mathbf{u}_a^\top\mathbf{u}_a}{2}.
\end{aligned}
$$

$$
\begin{aligned}
P(\nu_v^2|...) &\propto \mathcal{IG}(\nu_v^2|e, f) \times \mathcal{N}(\mathbf{z}_v|0, \nu_v^2\mathbf{I}), \\
&= \mathcal{IG}(\nu_v^2|e', f') \\
e' &= \frac{2e + P}{2} \\
f' &= \frac{2f + \mathbf{z}_v^\top\mathbf{z}_v}{2}.
\end{aligned}
$$

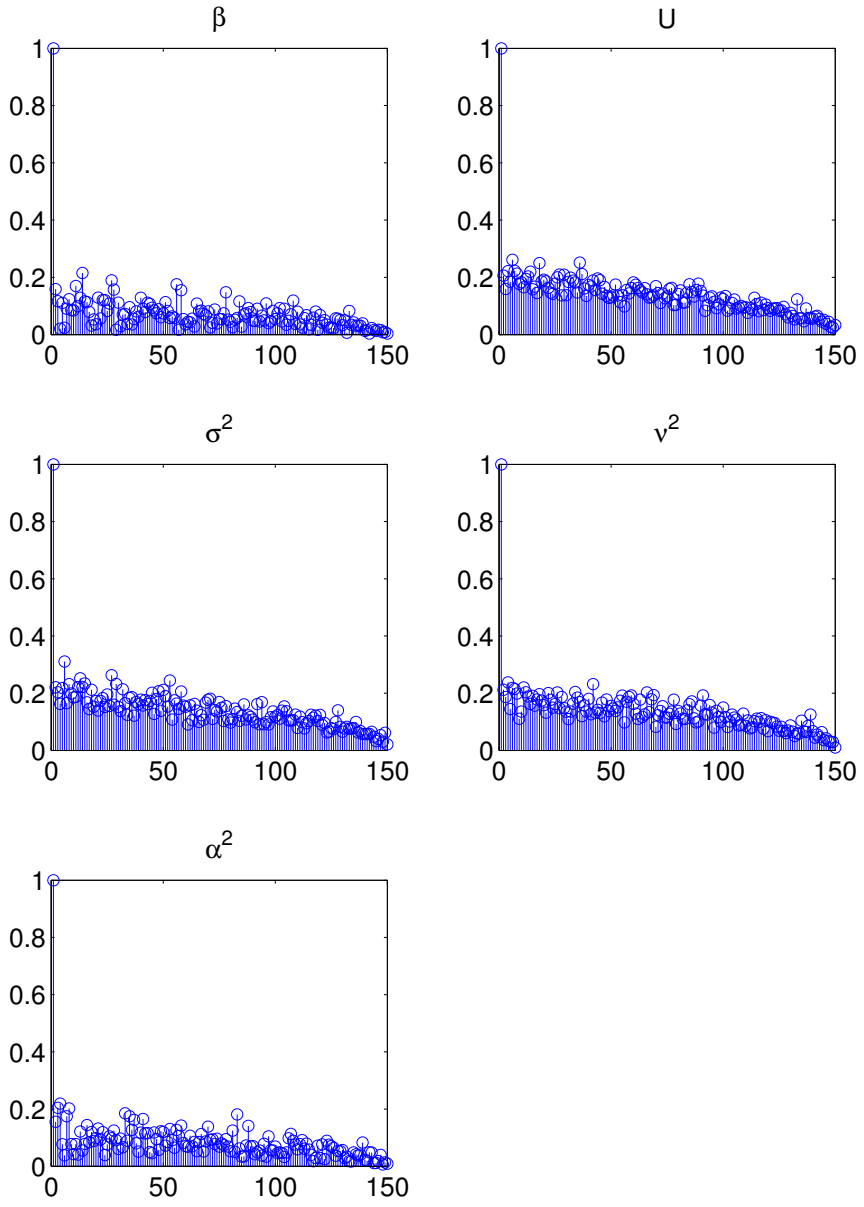


Figure A.1: Maximum autocorrelation plots after burn-in and thinning. We used a thinning of 10 and a burn-in of 100 resulting in 150 samples.

A.2 The Marginal Prior of the SAE Model

The small-area model has a Gaussian prior distribution $\mathbf{z}_v|\nu_v^2 \sim \mathcal{N}(0, \nu_v^2\mathbf{I})$ for the regression coefficients specific to voxel v. The voxel-specific variance has an inverse gamma prior distribution, where $\nu_v^2 \sim \mathcal{IG}(e, f)$. As mentioned in section 3.5.5, this implies that the *marginal* prior distribution of $\mathbf{z}_v$ is a scaled t distribution, where $\mathbf{z}_v/(f/e) \sim t_{2e}$ (see Gelman et al. 33, section 3.3). This t distribution approaches a Gaussian rather quickly as the number of degrees of freedom grows. Figure A.2 shows the density obtained from 10^4 draws from the hierarchical prior when $e = 3$ (so that there are 6 degrees of freedom), along with the theoretical t distribution, and the approximating Gaussian.

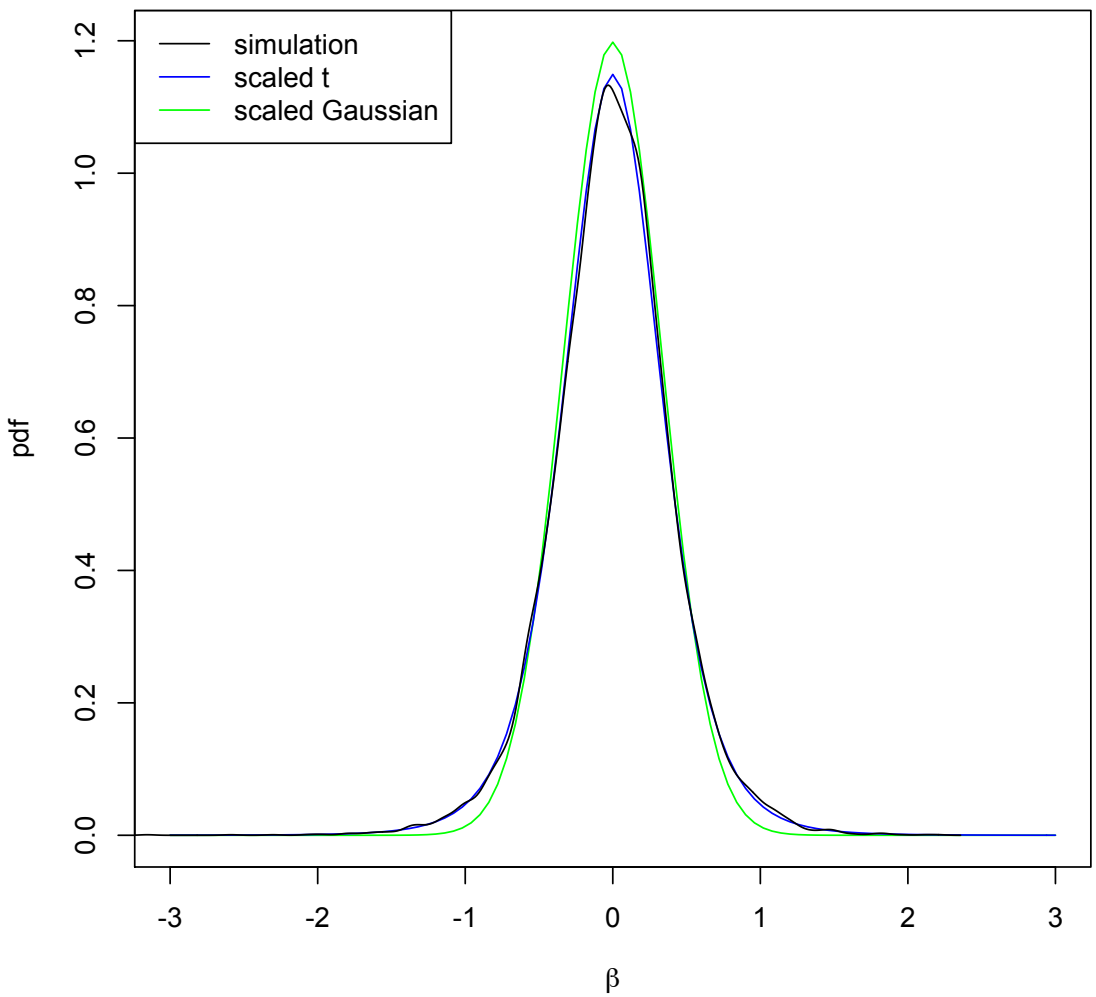


Figure A.2: The SAE model's marginal prior distribution for regression coefficients, when the hyper-parameter $e = 3$. The black line shows 10^4 draws from the hierarchical prior, the blue the theoretical t distribution (with 6 degrees of freedom), and the green the matching Gaussian.

A.3 Model Checking

The most important assumption of our models is the linearity of expected voxel activity as a function of stimulus features. If this holds, actual and predicted activities should themselves be linearly related. Figure A.3 shows that this holds tolerably. In the absence of a neurobiologically-grounded alternative, or enough data to make nonparametric estimates practical, we thus stay with the reasonable, and computationally cheap, linear model.

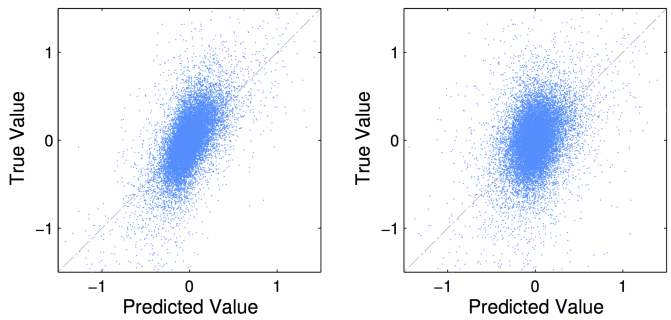


Figure A.3: Scatter plots of true activity versus predicted activity using ridge regression for in-sample data (left) and out-of-sample data (right) for 1000 voxels picked at random from the set of voxels with good classification accuracy (greater than 60%).

With the HB model, one must both check the behavior of the posterior distribution (as approximated by the output of the Gibbs sampler), and check the prior itself. The Gibbs sampler showed little change in either parameter estimates or predictive performance when varying the hyper-parameters α and β over an order of magnitude. Posterior predictive simulations (Appendix A.3.1) indicated that if the SAE model was well specified it should out-predict ridge. Since this is not the case with the data, this suggests model misspecification.

A.3.1 Simulation of the SAE Model

App. A.2 suggest that ridge regression and the SAE model should lead to very similar parameter estimates and hence to similar predictions. However, in our data we found the two to be virtually indistinguishable. Here, we show that this should *not* be the case if the SAE model were properly specified, and investigate what sort of mis-specification might account for it.

In these simulations, we simulate from the small-area model of section 3.5.2, drawing parameters from the prior. For speed and simplicity, we limit ourselves to 500 voxels, divided into 5 ROIs of 100 voxels each. The stimuli x were fixed to those employed in E2, and values of $y_v(t)$ draw conditional on x and the randomly-generated parameters. The surrogate values from this simulation were then fit to three models: ridge regression; the small-area model with the correct assignment of voxels to ROIs; and the small-area model with 100 voxels assigned to 5 ROIs at random.

Figures A.4 and A.5 show that the properly-specified small-area model has a modest, but systematic and significant, advantage in forward prediction over ridge regression. This disappears, however, when the small-area model is mis-specified because it gets the assignment of voxels to

areas wrong. As this predictive equivalence is more or less what we see in the data from both experiments, the left-hand panels of the figures lets us conclude that that the HB model is mis-specified *somehow*. The right-hand panels suggest, but do not prove, that the mis-specification arises from using the wrong division of voxels into regions.

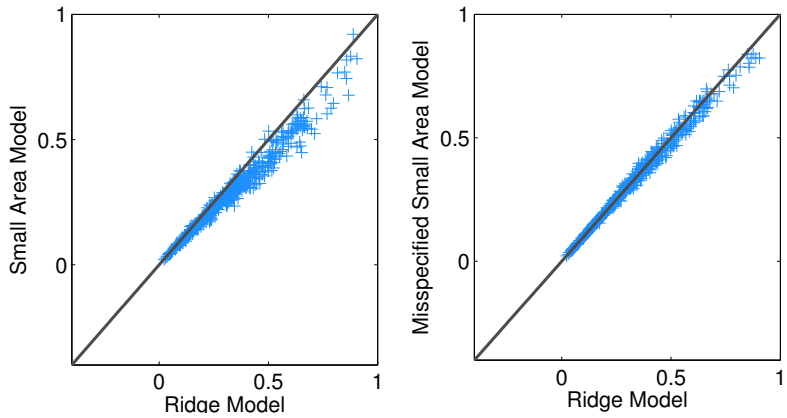


Figure A.4: Voxel-wise RSS for the small-area model (vertical axis) versus that for ridge regression (horizontal), fit to simulations of the small-area model, with the assignment of voxels to ROIs being either correct (left) or incorrect (right). The small area estimates outperforms the ridge estimates when the true underlying model is a small area model and the small area estimator is correctly specified.

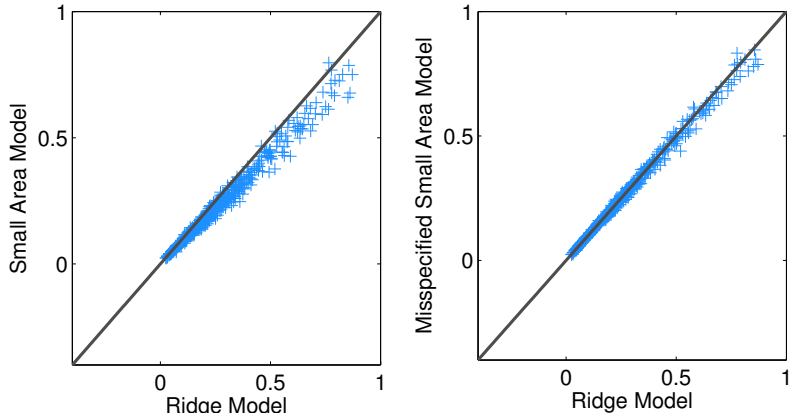


Figure A.5: Same as in Figure A.4, but increasing the extend to which shared area parameters vary between areas. Again, the small area estimates outperforms the ridge estimates when the true underlying model is a small area model and the small area estimator is correctly specified.

A.4 Regularization Reduces Variability

Figure A.6 shows how regularization, either by the ridge penalty or the prior of the SAE model, reduces standard errors in parameter estimates, compared to OLS. This indicates that the problem is one of small-area estimation in the technical sense, and needs some form of regularization.

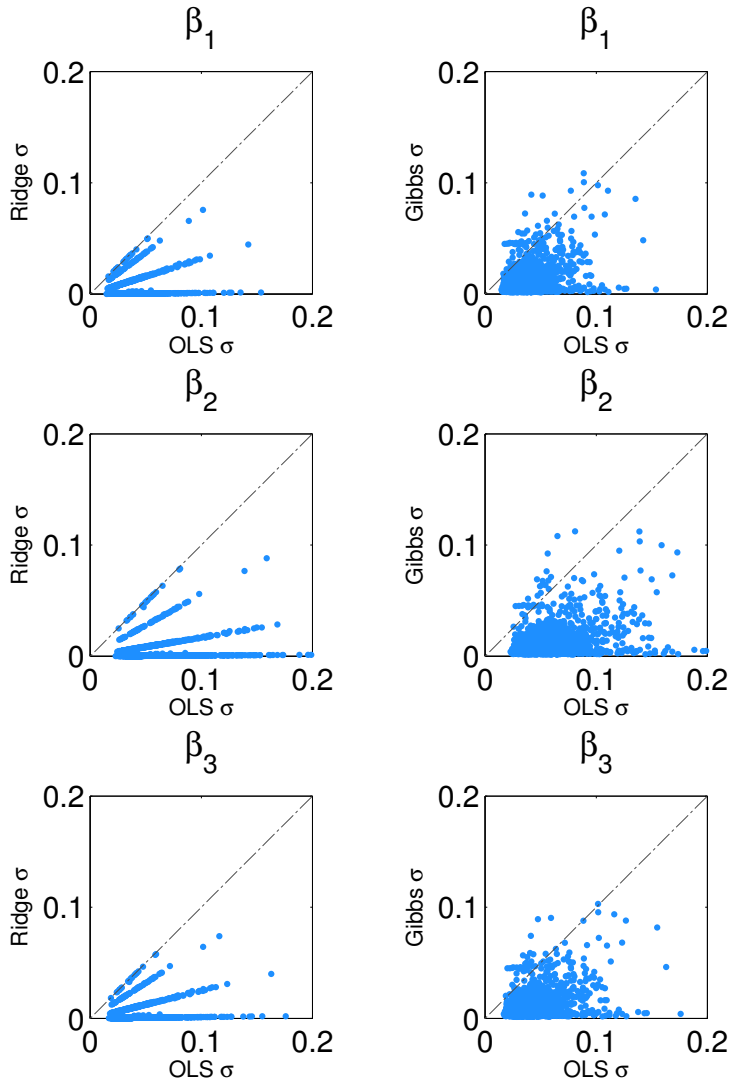


Figure A.6: Standard errors of voxel-wise ridge-regression estimates (left column) and of SAEs (right), versus the standard errors of direct OLS estimates. Regularization (either with Ridge or SAE) reduces the standard errors of the estimates, indicating that the parameters need some form of regularization.

A.5 Replication of the experiment

We repeat the analysis described in chapter 3 with the same data and experimental design from chapter 4. We use non-smoothed fMRI data, and the visual features (average word length and standard deviation of word length) as an IFS. We repeat the analysis with OLS, ridge regression, elastic net and the small area bayesian model. We call this experiment E2. We call the trial based experiment used in chapter 3 from [79] E1. E1 and E2 extremely different. Common findings about the properties and performance of statistical methods across such different settings are very unlikely to be artifacts of a *particular* experiment. We present here the figures that are analogous to figures 3.5, 3.6, 3.7 and 3.9 and that show remarkably similar patterns.

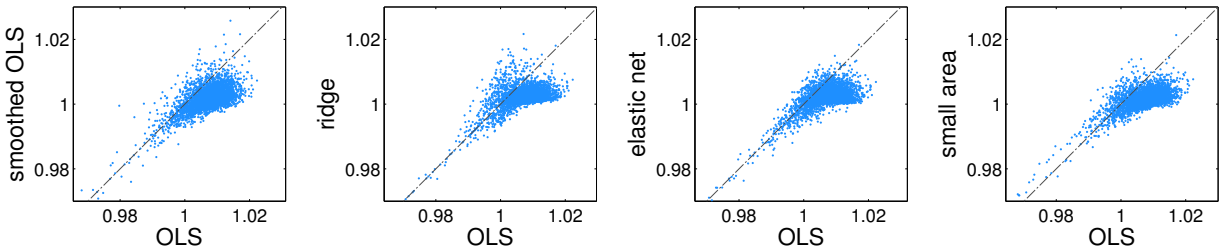


Figure A.7: Effect of regularization on out-of-sample normalized RSS (RSS/σ^2). For each of the plots, the OLS RSS/σ^2 (horizontal axis) is contrasted with the modified RSS/σ^2 after OLS smoothing for ridge, elastic net or small area shrinkage (vertical axis). The four methods result in smaller RSS/σ^2 on average. Furthermore, for all the methods, the predicted activity in the bad voxels (i.e. voxels where RSS/σ^2 is larger than 1) is pushed toward zero. This is visible by the RSS/σ^2 values being reduced toward 1. In other words, shrinkage and smoothing are forcing the estimated parameters to be almost zero if the voxel is noisy and there is nothing that can be predicted.

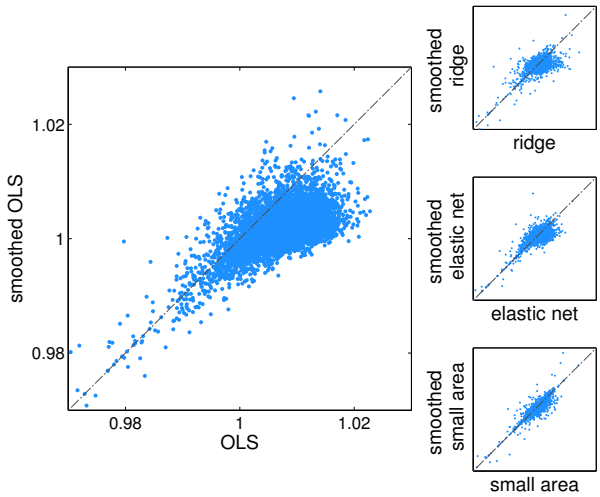


Figure A.8: Normalized RSS for unsmoothed and smoothed estimators. The larger panels show voxel-wise normalized residuals (RSS/σ^2) for OLS before smoothing (horizontal axis) and after (vertical), showing the value of spatial smoothing for forward inference. The smaller panels consist of the same comparison for ridge regression (top), the elastic net (middle) and the small-area model (bottom), showing that combining smoothing and shrinkage is if anything worse than shrinkage alone. The axes for the smaller panels have been omitted for clarity: they correspond to the larger panels axes.

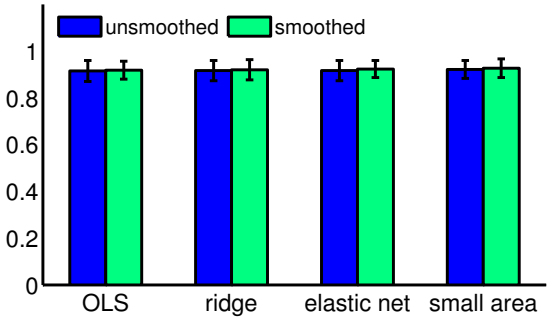
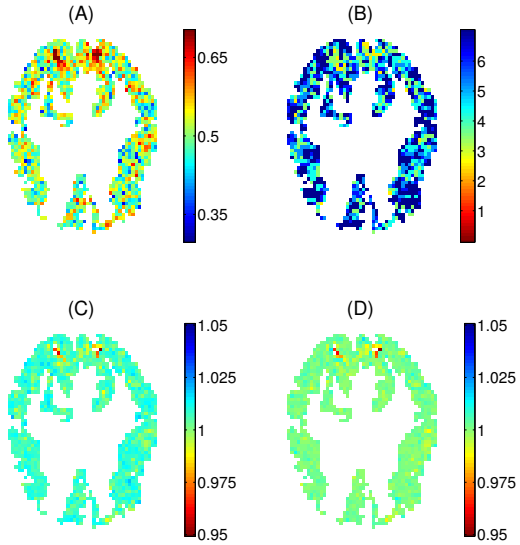
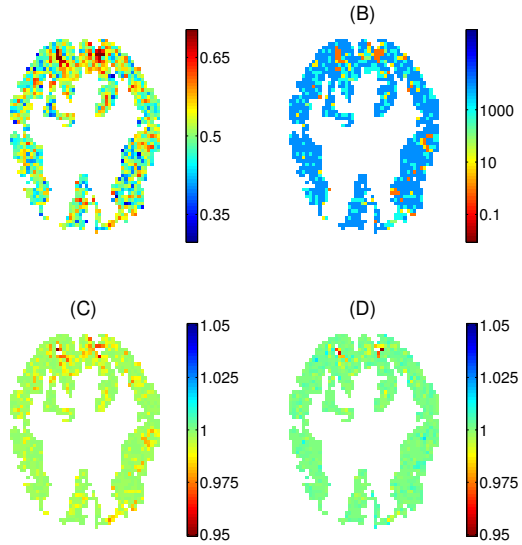


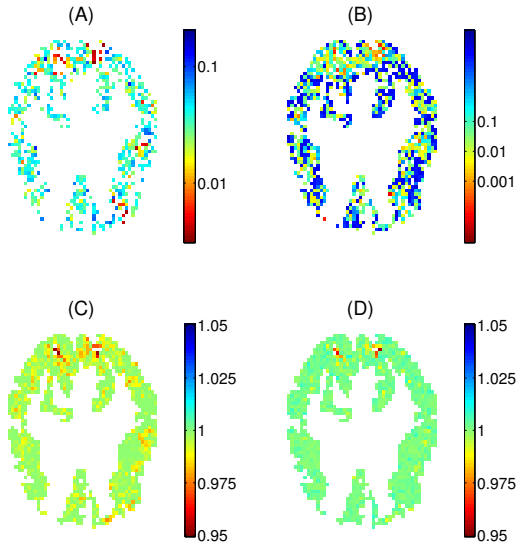
Figure A.9: Whole-brain classification accuracy, averaging over subjects, for all combinations of estimators and smoothing. Regularization choice or the presence or absence of smoothing don't affect whole-brain classification accuracy.



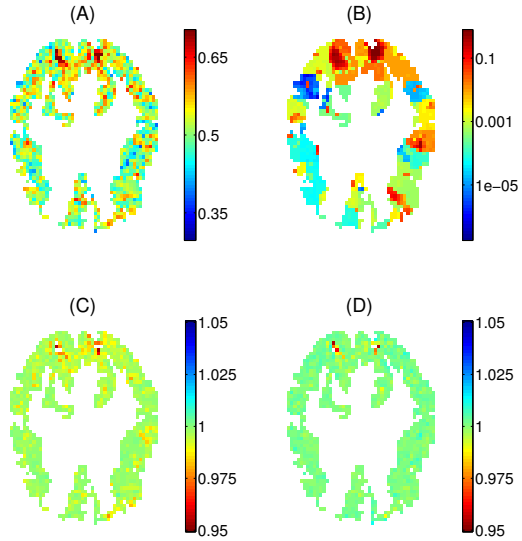
(a) **OLS:** A, classification accuracy; B, smoothing radius; C, D, normalized out-of-sample RSS pre- and post- smoothing



(b) **Ridge:** A, classification accuracy; B, λ parameter; C, D, normalized RSS in- and out-of- sample



(c) **Elastic Net:** A, λ_1 (lasso penalty); B, λ_2 (ridge penalty); C, D, normalized RSS in- and out-of- sample



(d) **Small Area:** A, classification accuracy; B, posterior mean variance of $\mathbf{z}_v$; C, D, normalized RSS in- and out-of- sample

Figure A.10: Voxel-wise results for each method along one horizontal brain slice. Color schemes are flipped so that red always represents "good" and blue, "bad". Note the similar patterns of classification accuracy in plots a-A, b-A and d-A. Also note how predictive performance (sub-figures A and D) is inversely related to the degree of regularization in every case, whether that is the smoothing radius for OLS (a), the λ penalty for ridge (b), the λ_1 and λ_2 penalties for the elastic net (c), or the small area model (d), where low regularization corresponds to a high variance parameter, i.e., good voxels are allowed to pick their parameters freely. (For the elastic net, good voxels have more lasso-like penalties, as they are voxels sensitive to *some* of the stimulus features.) Smoothing acts as a regularizer for OLS, as seen by the reduced prediction in the bad voxels from subfigure a-C to subfigure a-D. Finally, see that in many cases the in- and out-of- sample errors for "good" voxels are nearly the same.

A.6 What is the effect of smoothing and regularization?

To see the effect of smoothing and regularization on OLS we compared the held out normalized RSS before and after smoothing, and with and without regularization in section 3.5.5. Here, we show the effect on the single voxel accuracies for both experiments. As seen in section 3.5.5, smoothing seems to reduce the forward prediction error in the worst voxels, but its effects on classification accuracy are at best ambiguous (Figure A.12). Much the same is true of shrinkage (Figure A.11).

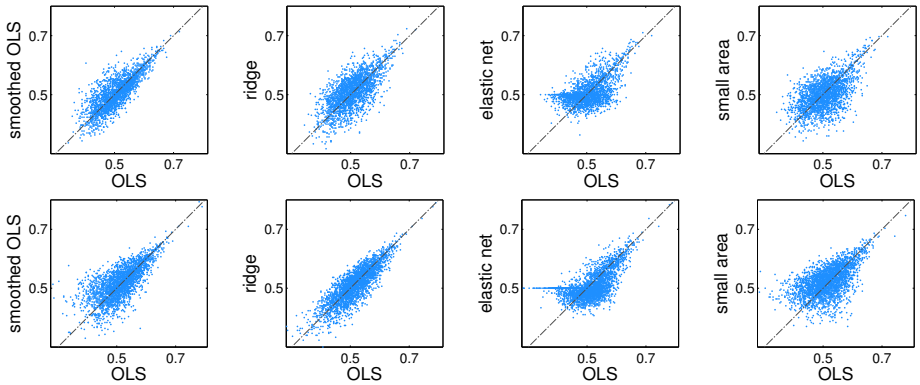


Figure A.11: Effect of smoothing or shrinkage on voxel-wise classification accuracy for E1 (top) and E2 (bottom). For each of the plots, the OLS accuracy (horizontal axis) is contrasted with the modified accuracy after OLS smoothing or ridge, elastic net or small area shrinkage (vertical axis). For both experiments, is it hard to discern any systematic effect of smoothing or shrinkage in this reverse inference task.

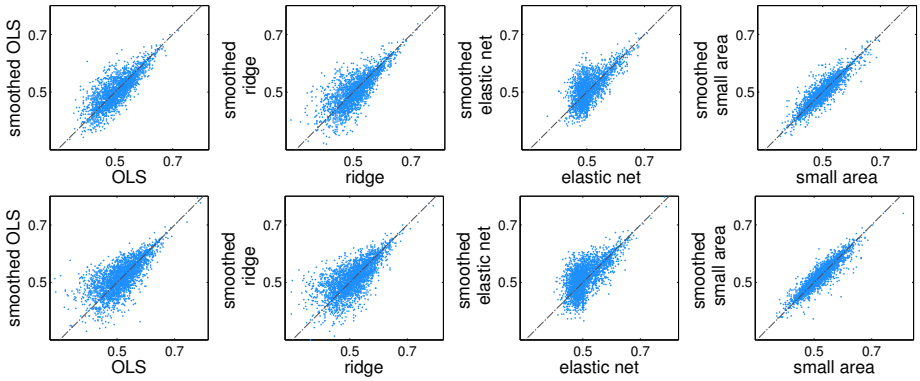


Figure A.12: Effect of smoothing on voxel-wise classification accuracy for E1 (top) and E2 (bottom). For each of the plots, the unsmoothed OLS, ridge, elastic net or small area estimators (horizontal axis) are contrasted with their smoothed version (vertical axis). For both experiments, is it hard to discern any systematic effect of smoothing or shrinkage in this reverse inference task.

A.7 Full Brain results

A.7.1 Experiment 1 (E1)

This section has 4 plots, which summarize our findings for the four methods of OLS, ridge regression, elastic net and the small area model for the first experiment (E1) described in 3. Color schemes are flipped so that red always represents "good" and blue, "bad". The images are best viewed on a computer screen with high resolution.

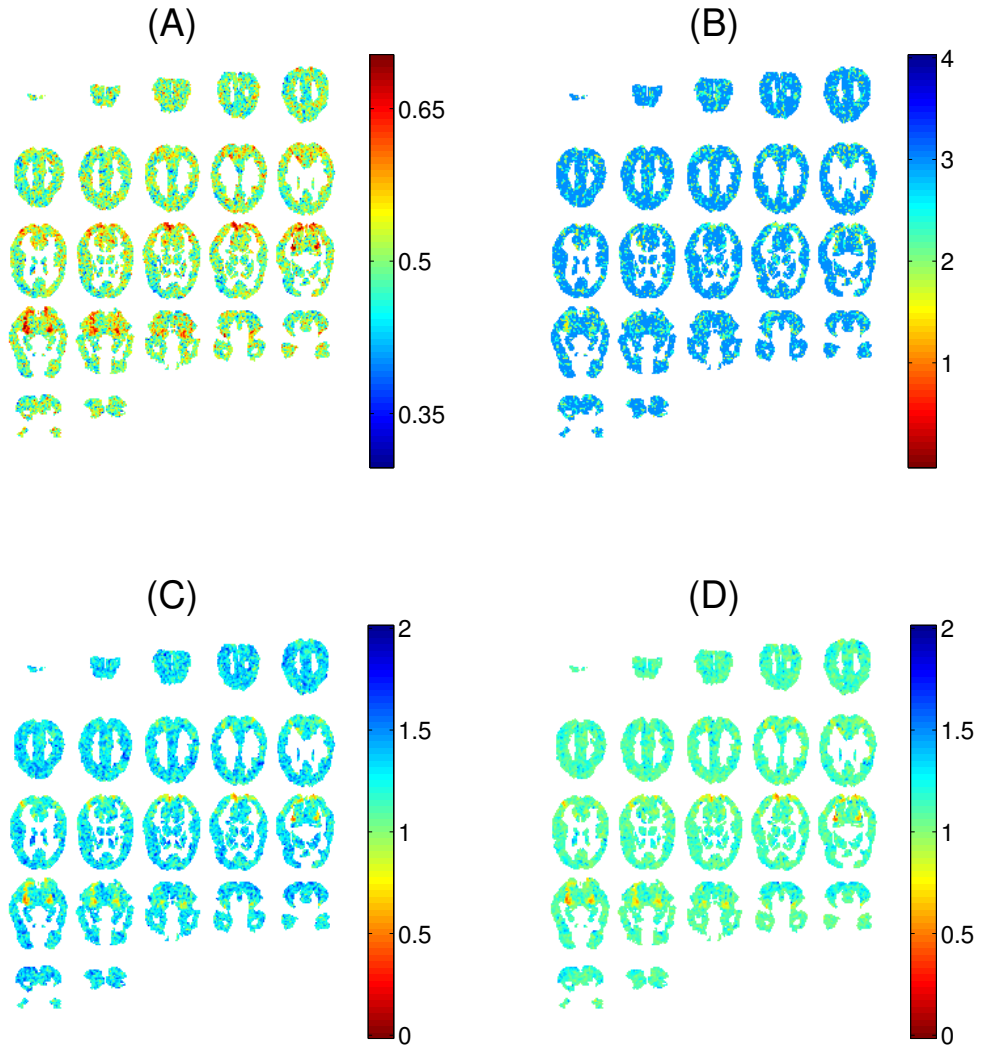


Figure A.13: E1: OLS classification accuracy (A), smoothing radius (B) and normalized out-of-sample RSS before (C) and after smoothing (D). Note how predictive performance (sub-figures A and D) is inversely related to the degree of regularization, which in this case is the is the smoothing radius for OLS. Note the improvement of the normalized out-of-sample RSS in the bad voxels after smoothing (indicated by less blue voxels in subplot D than subplot C).

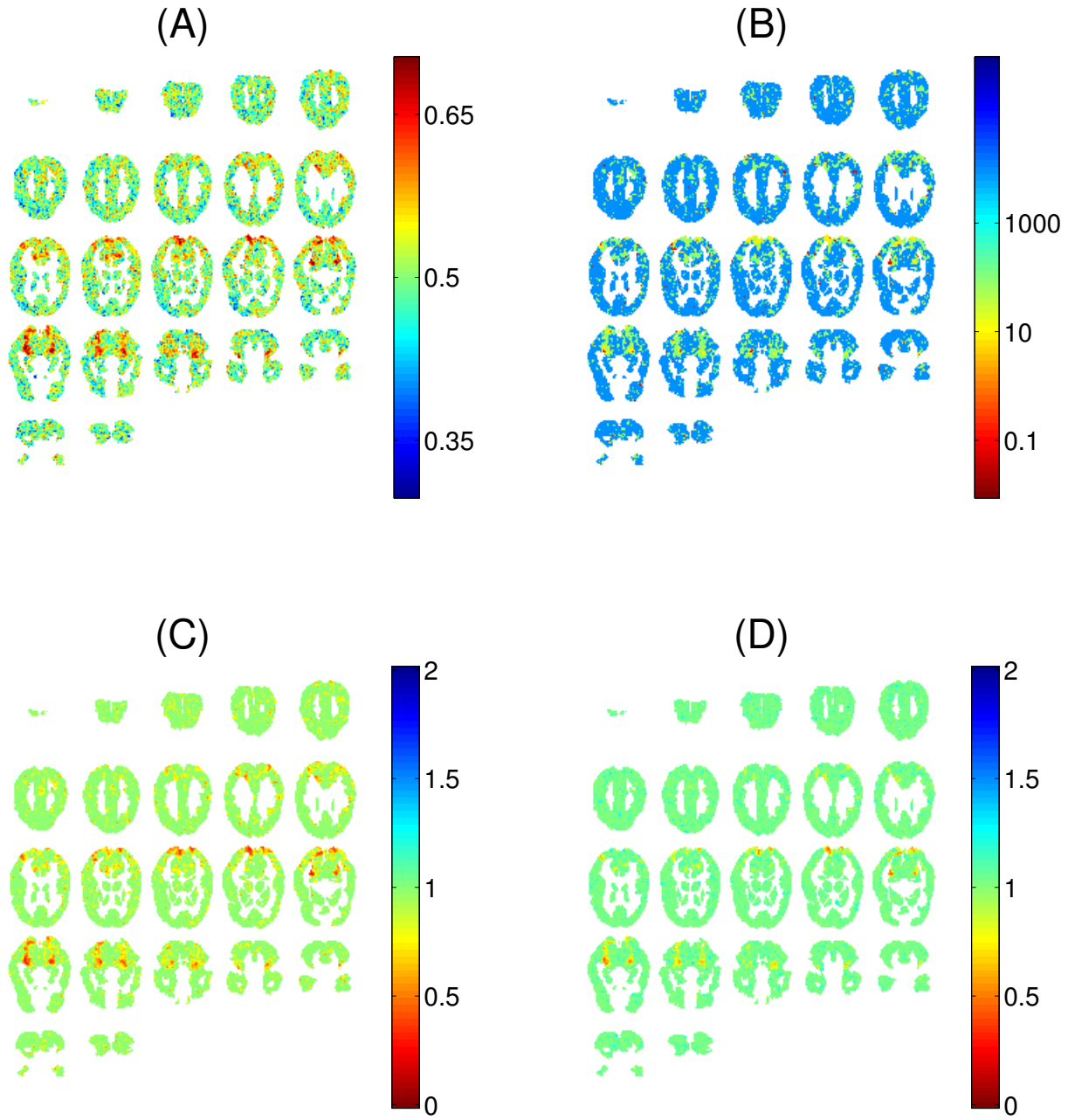


Figure A.14: E1: ridge regression classification accuracy (A), λ (B) and normalized RSS in (C) and out of sample (D). Note how predictive performance (sub-figures A and D) is inversely related to the degree of regularization which in this case is the λ penalty for ridge. Low regularization corresponds to accurate voxels, i.e., these "good" voxels are allowed to pick their parameters more freely. The in- and out-of- sample errors for "good" voxels are nearly the same.

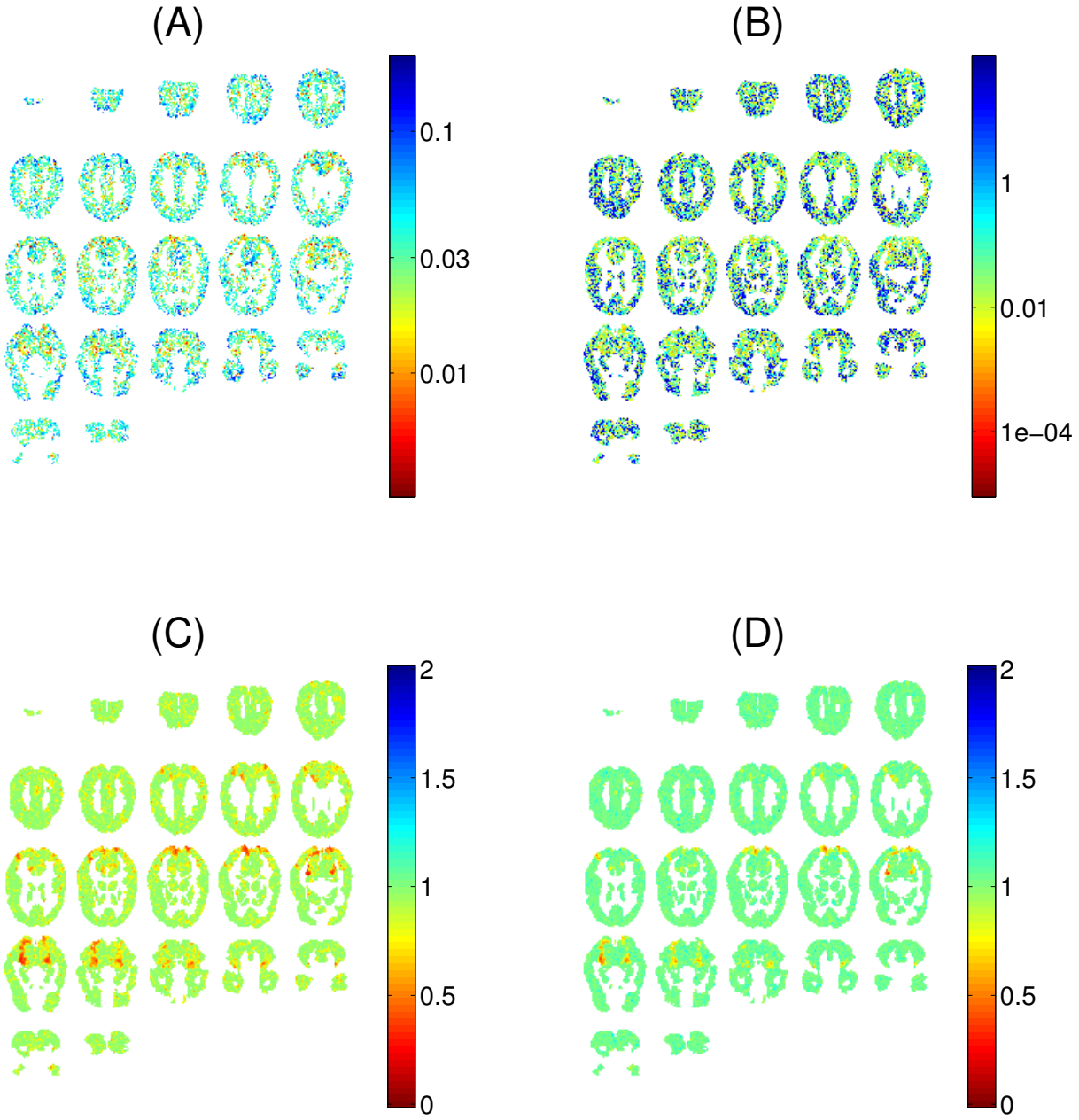


Figure A.15: E1: Elastic net λ_1 (lasso penalty) (A), λ_2 (ridge penalty) (B) and normalized RSS in (C) and out of sample (D). Note how predictive performance (sub-figure D) is inversely related to the degree of regularization, which in this case is the λ_1 and λ_2 penalties for the elastic net. Good voxels have more lasso-like penalties, as they are voxels sensitive to *some* of the stimulus features. The in- and out-of- sample errors for "good" voxels are nearly the same.

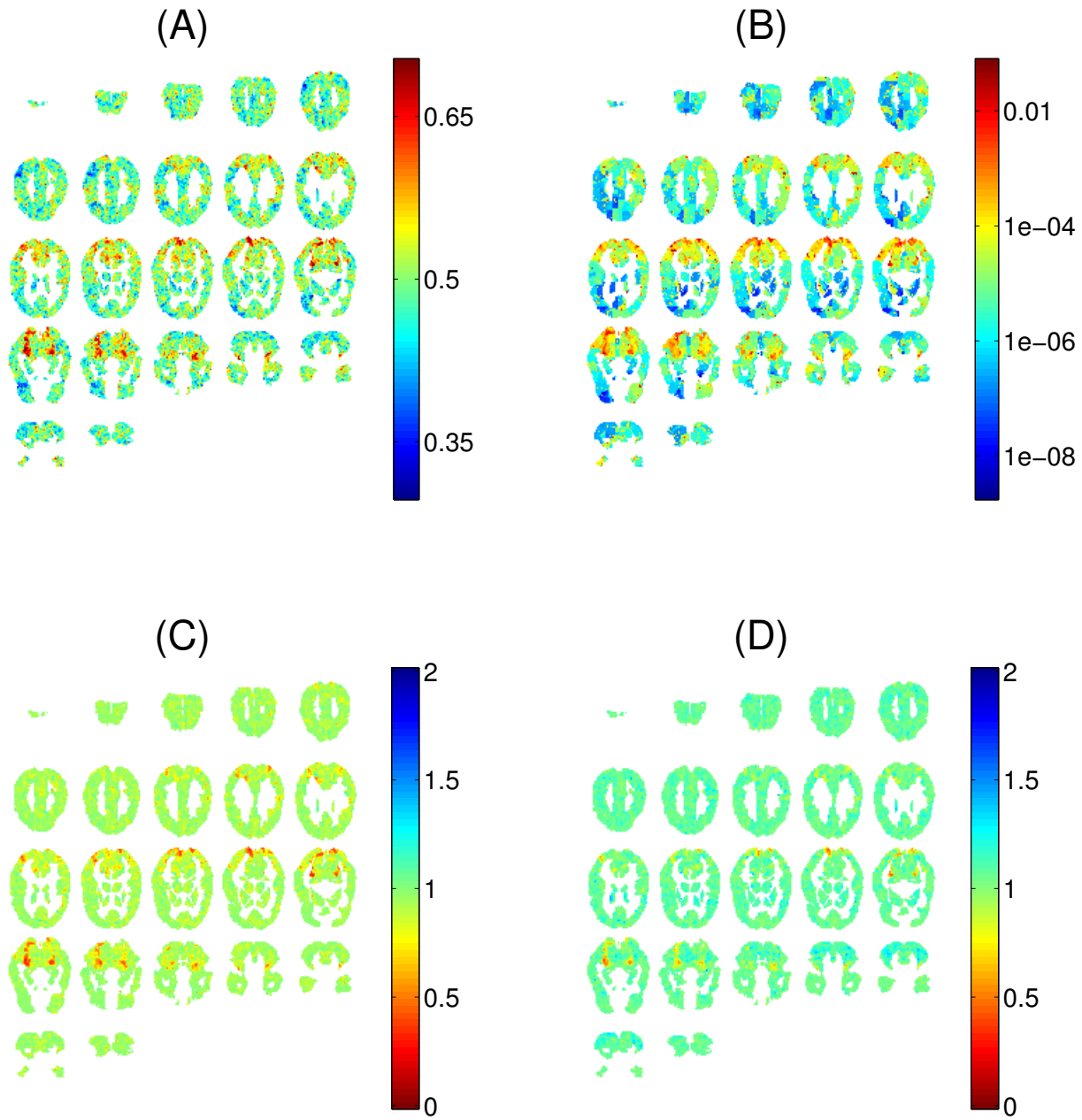


Figure A.16: E1: small-area model classification accuracy (A), posterior mean of the variance of β_v per voxel (B) and normalized RSS in (C) and out of sample (D). Note how predictive performance (sub-figures A and D) is inversely related to the degree of regularization of the small area model: low regularization corresponds to a high variance parameter, i.e., good voxels are allowed to pick their parameters freely. The in- and out-of- sample errors for "good" voxels are nearly the same.

A.7.2 Experiment 2 (E2)

This section has 4 plots, which summarize our findings for the four methods of OLS, ridge regression, elastic net and the small area model for the second experiment (E2) described in chapter 3. Color schemes are flipped so that red always represents "good" and blue, "bad". The images are best viewed on a computer screen with high resolution.

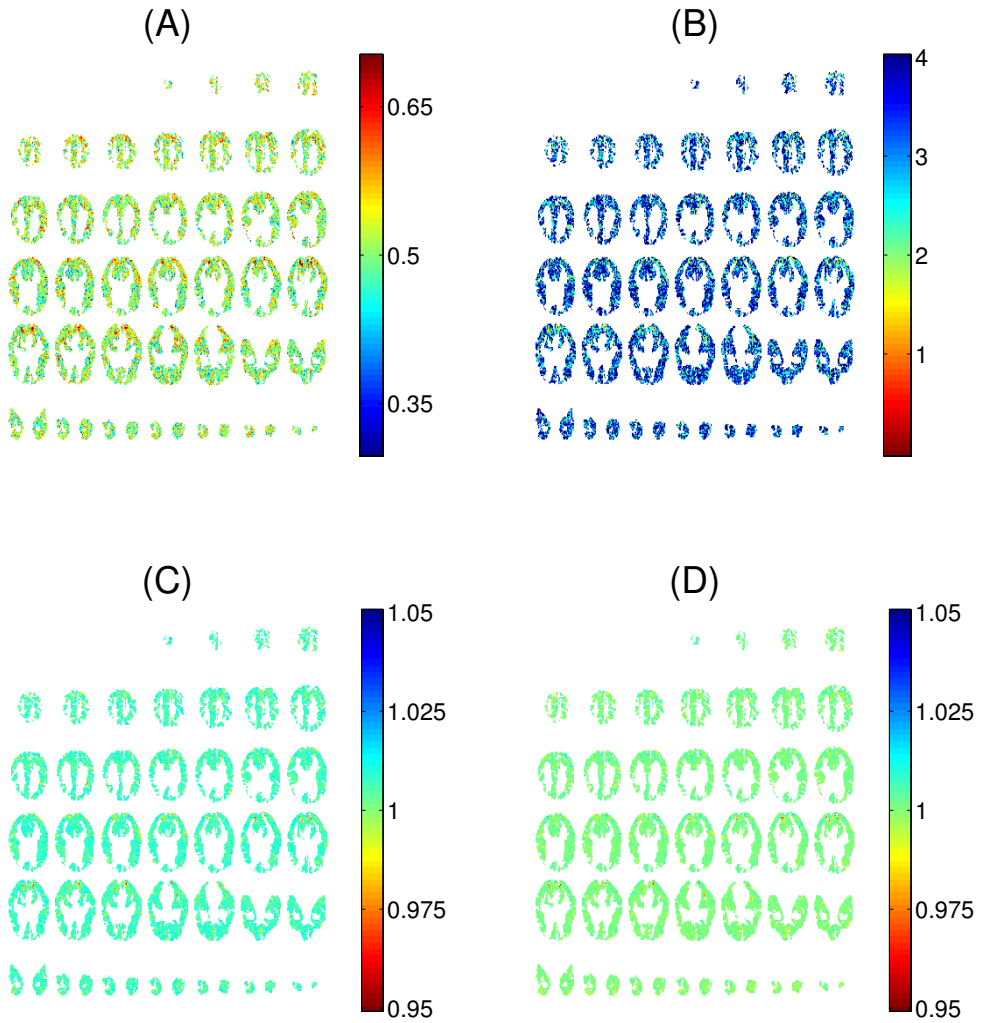


Figure A.17: E2: OLS classification accuracy (A), smoothing radius (B) and normalized out-of-sample RSS before (C) and after smoothing (D). Note how predictive performance (sub-figures A and D) is inversely related to the degree of regularization, which in this case is the is the smoothing radius for OLS. Note the improvement of the normalized out-of-sample RSS in the bad voxels after smoothing (indicated by less blue voxels in subplot D than subplot C).

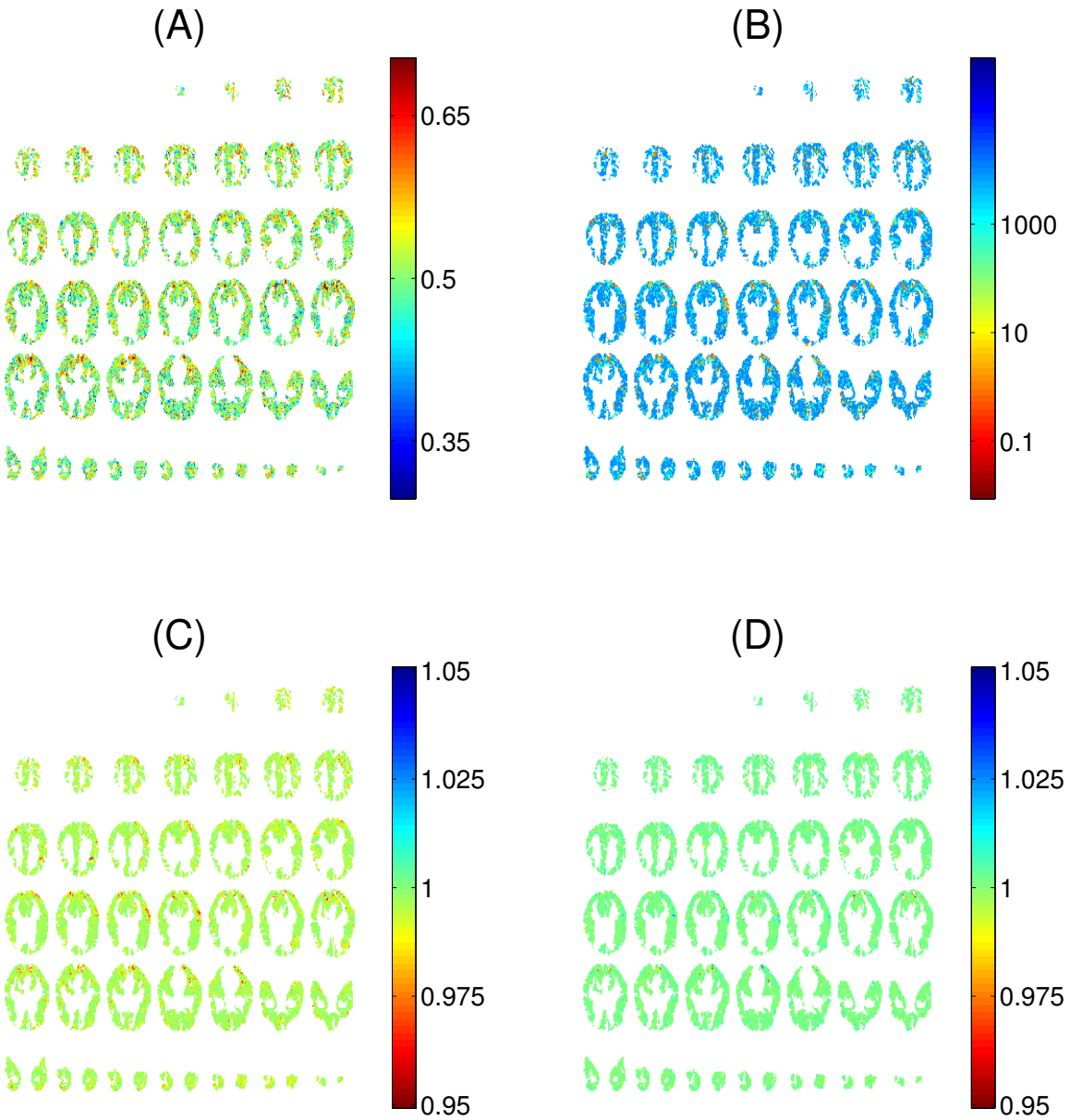


Figure A.18: E2: ridge regression classification accuracy (A), λ (B) and normalized RSS in (C) and out of sample (D). Note how predictive performance (sub-figures A and D) is inversely related to the degree of regularization which in this case is the λ penalty for ridge. Low regularization corresponds to accurate voxels, i.e., these "good" voxels are allowed to pick their parameters more freely. The in- and out-of- sample errors for "good" voxels are nearly the same.

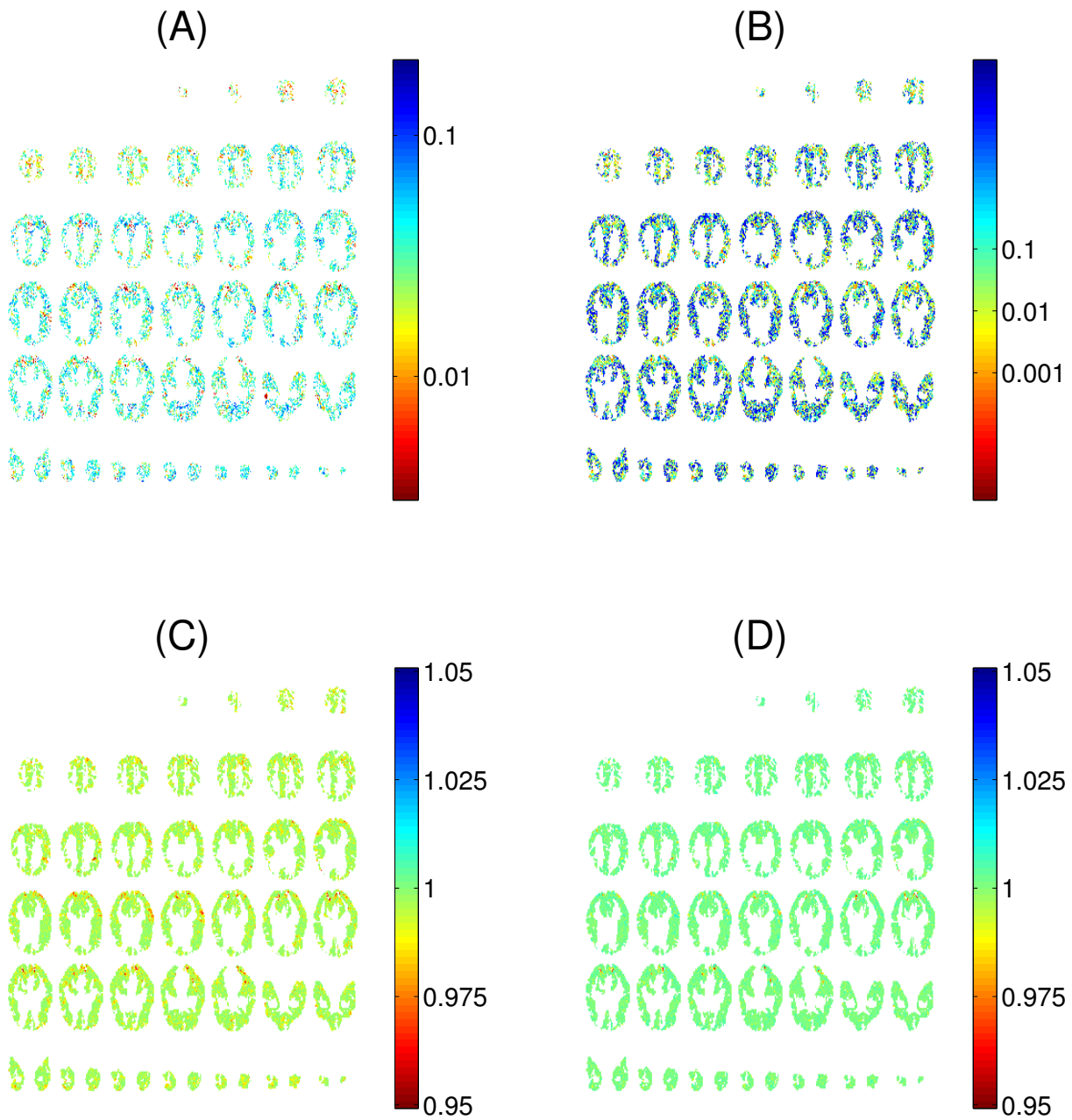


Figure A.19: E2: Elastic net λ_1 (lasso penalty) (A), λ_2 (ridge penalty) (B) and normalized RSS in (C) and out of sample (D). Note how predictive performance (sub-figure D) is inversely related to the degree of regularization, which in this case is the λ_1 and λ_2 penalties for the elastic net. Good voxels have more lasso-like penalties, as they are voxels sensitive to *some* of the stimulus features. The in- and out-of- sample errors for "good" voxels are nearly the same.

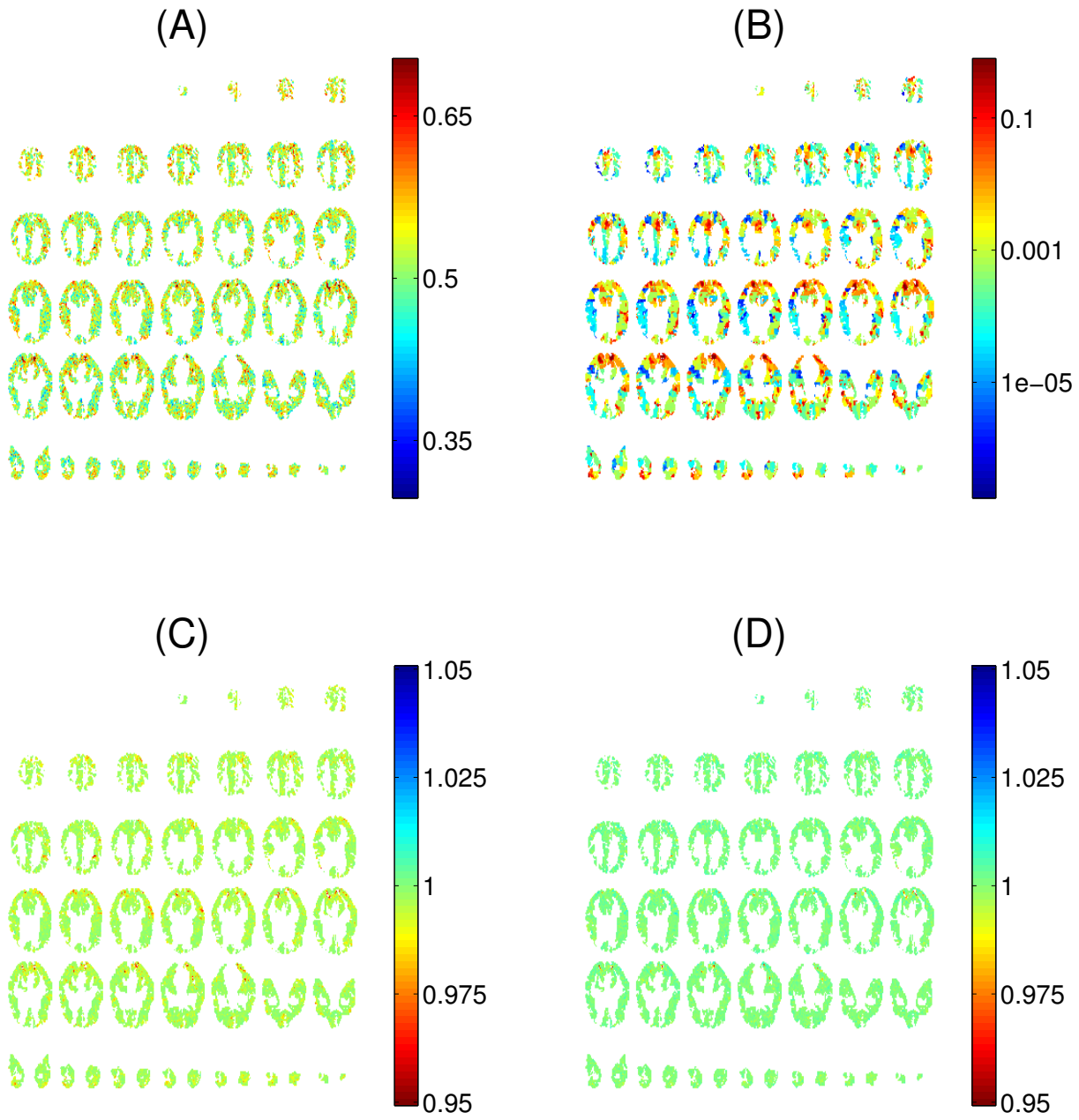


Figure A.20: E2: small-area model classification accuracy (A), posterior mean of the variance of β_v per voxel (B) and normalized RSS in (C) and out of sample (D). Note how predictive performance (sub-figures A and D) is inversely related to the degree of regularization of the small area model: low regularization corresponds to a high variance parameter, i.e., good voxels are allowed to pick their parameters freely. The in- and out-of- sample errors for "good" voxels are nearly the same.

Appendix B

Examples of Estimated Parameters

After learning the set of parameters, we look at the four points we learned for a feature j at a voxel v and examine their relative shape. We find that the responses learned are very noisy. However when only looking at the average response for a given feature type at the regions that represent this feature type (we obtain these regions via the classification task explained in detail in the next section), we end up with 4 points that can usually be fitted on a concave waveform that resemble the characteristic shape of the hemodynamic response. We present the average waveforms we learned in Fig. 6. It should be noted that these plots are the averages by feature set, for one of the subjects, of parameters learned across the voxels whose accuracy is in the top 95% percentile, and therefore they are only provided as an illustration.

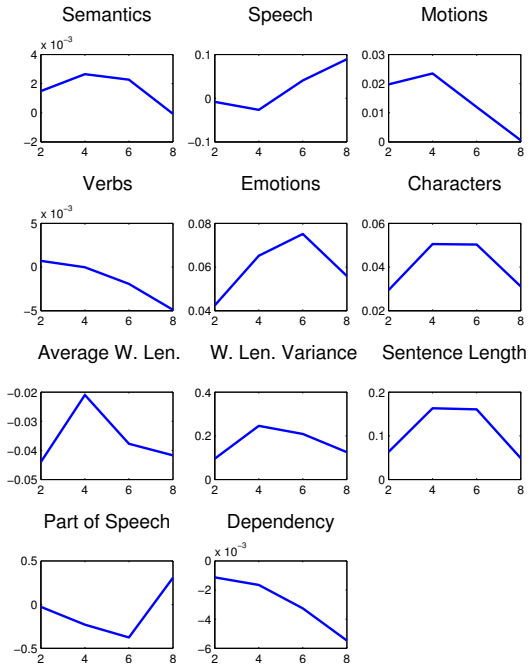


Figure B.1: Global averages of the parameters learned for each feature type.

Appendix C

Additional Results for Chapter 4

We present here the 3D map we obtain for syntactic features exclusively, divided into the contribution from our three types of syntactic features: sentence length, part of speech and dependency roles.

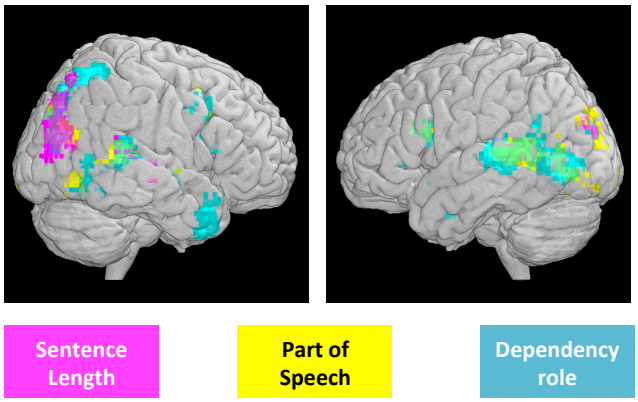


Figure C.1: Results obtained by our generative model for different syntax features, showing where sentence length, part of speech, and dependency roles are encoded by neural activity. Each voxel location represents the classification when using a cube of $5 \times 5 \times 5$ voxel coordinates, centered at that location, such that the union of voxels from all subjects whose coordinates are in that cube are used. Voxel locations are colored according to the feature set that can be used to yield significantly higher than chance accuracy.

We have also ran the entire experiment with the same setup, using however the data without spatial smoothing. The results vary to a considerable degree in the boundaries of each region, while the main location of each feature representation stays the same. Figures C.2 and C.3 show the resulting maps.

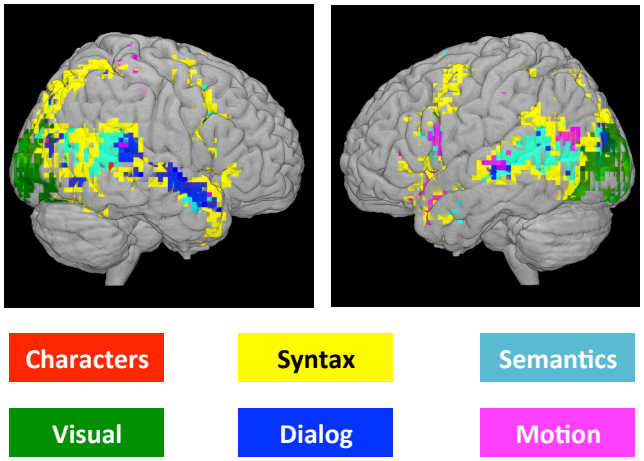


Figure C.2: Same as figure 4 with non-smoothed data (at FDR $\alpha = 0.01$).

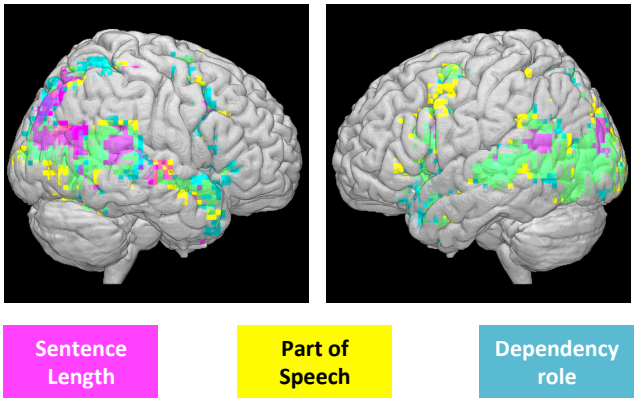


Figure C.3: Same as figure C.1 with non-smoothed data.

Our results do not only depend on processing methods, but they also require the significance thresholding of different classification tasks which might not be of equal difficulty. For instance, different features might lead to high or low classification because of the statistical properties of the features and not the way they are represented in the brain. We present below the comparison of the whole brain classification when different types of features are used. We compare these accuracies with the entropy of each feature set. We want to see if the difference in classification accuracy is due to differences in the entropy of each feature: it is harder to learn a model with features that change rarely in a story (low entropy), than it is to learn a model with features that occur very frequently. In our feature creation phase, we did explicitly exclude features with low entropy (for example, the location of scenes didn't vary much and we didn't include it). However, the features we did keep still vary in their frequency and we wanted to compare their entropies to their accuracies.

For each feature set we compute the entropy of each feature, and then use the maximum entropy. The results are shown in the first row of tables C.1 and C.2. In the following rows, we show classification accuracy by feature set. For the smoothed data, the accuracy was initially low

and was boosted by voxel selection as explained in chapter 3.

	NNSE	Average WL	Variance WL	Sentence Length
entropy	1.84	4.01	5.22	5.46
accuracy (smoothed)	0.58	0.69	0.57	0.53
boosted accuracy (smoothed)	0.63	0.87	0.80	0.62
accuracy (unsmoothed)	0.75	0.71	0.71	0.67

Table C.1: Non-binary features.

	speak	move	emotions	verbs	characters	POS	dependency
entropy	0.69	0.50	0.25	0.28	0.28	0.62	0.78
accuracy (smoothed)	0.55	0.51	0.50	0.51	0.56	0.61	0.62
boosted accuracy (smoothed)	0.66	0.61	0.50	0.65	0.52	0.71	0.71
accuracy (unsmoothed)	0.69	0.61	0.48	0.65	0.56	0.84	0.83

Table C.2: Binary features.

There seems to be a modest relationship between the entropy of the features and how accurate classification is, in which feature sets with higher entropy lead to a higher accuracy. There might be other factors also affecting how easy the classification with different feature sets are. To avoid comparing the results of classification tasks that vary in difficulty, and as a way of leveling the playing field, we plot in Fig. C.4 the top 1000 voxels when using each of the feature sets. The voxels that are colored do not therefore necessarily have a higher than chance classification accuracy.

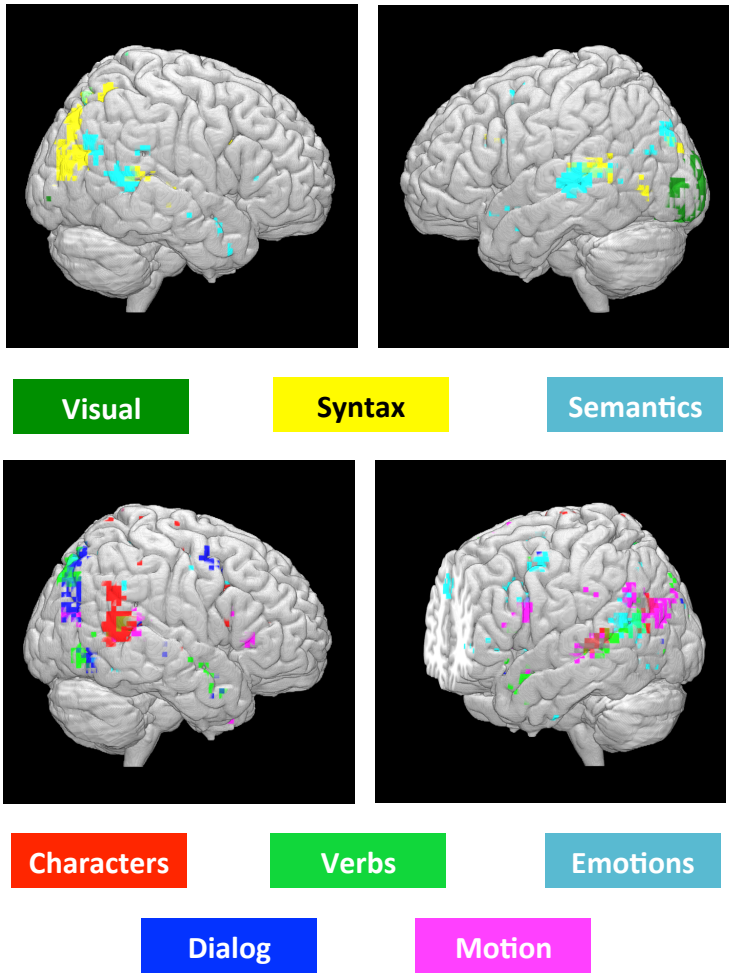


Figure C.4: Top 1000 voxels for each feature type (smoothed data). Instead of picking the significantly higher than chance voxels, we chose to color the 1000 voxels with the highest (normalized) accuracy for each feature type. The accuracies were normalized using the empirical null distribution as explained in Appendix F. In the lower, right figure, the brain is sliced to reveal in the medial frontal cortex a cluster of voxels that in which emotions lead to relatively high accuracy.

Appendix D

Combining Subjects Spatially

Our concatenated Searchlight is not equivalent to spatial or cross-participant smoothing because, again, the voxels associated with each subject are treated independently. The only requirement is that the subjects are all normalized to the MNI space; we do not co-register the subjects and we learn the response of every voxel independently.

Assume we are interested in an area A that is distributed around a certain mean location (x, y, z) in all subjects. Then despite the subjects' anatomical variability and given an adequate model and an appropriate cube-size, the cube centered at (x, y, z) will contain in it the voxels from area A of all subjects. Running the classification at this cube should then hypothetically yield the best accuracy. This would be possible because, inside the cube, the voxels from all subjects are concatenated and they contribute independently to the Euclidean distance we compute in classification. The voxels' precise alignment is irrelevant at this step, it only matters that they are all taken into consideration. Therefore, this method identifies regions of a given size (in this case 15mm $\times$ 15mm $\times$ 15mm) in which the subjects are processing the same information. It avoids the problem usually encountered in averaging multiple subjects, which is that the only regions that are identified are the regions in which the subjects highly overlap. This problem is widely debated in the literature [21].

Furthermore, despite the linearity of the model, this approach does not yield the same results as spatially smoothing the data in the cubes, because we have a multivariate input (the different story features) and while nearby voxels might be processing the same type of information (e.g. story characters), they are hypothetically coding different instances of this information (e.g. different story characters) with different patterns of activity for each instance.

Appendix E

Nonparametric Independence Testing for Small Sample Sizes

This chapter deals with the problem of nonparametric independence testing, a fundamental decision-theoretic problem that asks if two arbitrary (possibly multivariate) random variables X, Y are independent or not, a question that comes up in many fields like causality and neuroscience. While quantities like correlation of X, Y only test for (univariate) linear independence, natural alternatives like mutual information of X, Y are hard to estimate due to a serious curse of dimensionality. A recent approach, avoiding both issues, estimates norms of an *operator* in Reproducing Kernel Hilbert Spaces (RKHSs). Our main contribution is strong empirical evidence that by employing *shrunk* operators when the sample size is small, one can attain an improvement in power at low false positive rates. We analyze the effects of Stein shrinkage on a popular test statistic called HSIC (Hilbert-Schmidt Independence Criterion). Our observations provide insights into two recently proposed shrinkage estimators, SCOSE and FCOSE - we prove that SCOSE is (essentially) the optimal linear shrinkage method for *estimating* the true operator; however, the non-linearly shrunk FCOSE usually achieves greater improvements in *test power*. This work is important for more powerful nonparametric detection of subtle nonlinear dependencies for small samples.

E.1 Stein shrinkage for improved power

The problem of *nonparametric* independence testing deals with ascertaining if two random variables are independent or not, making no parametric assumptions about their underlying distributions. Formally, given n samples (x_i, y_i) for $i \in \{1, ..., n\}$ where $x_i \in \mathbb{R}^p, y_i \in \mathbb{R}^q$, that are drawn from a joint distribution P_{XY} supported on $\mathcal{X} \times \mathcal{Y} \subseteq \mathbb{R}^{p+q}$, we want to decide between the *null* and *alternate* hypotheses

$$\mathcal{H}_0 : P_{XY} = P_X \times P_Y \text{ vs. } \mathcal{H}_1 : P_{XY} \neq P_X \times P_Y$$

where P_X, P_Y are the marginals of P_{XY} w.r.t. X, Y. A test is a function from the data to $\{0, 1\}$. Tests aim to have high power (probability of detecting dependence, when it exists) at a

prespecified allowable type-1 error rate α (probability of detecting dependence when there isn't any).

Independence testing is often a precursor to further analysis. Consider for instance conditional independence testing for inferring causality, say by the PC algorithm [115], whose first step is (unconditional) independence testing. It is also useful for scientific discovery like in neuroscience, to see if a stimulus X (say an image) is independent of the brain activity Y (say fMRI) in a relevant part of the brain. Since *detecting* nonlinear correlations is much easier than *estimating* a nonparametric regression function (of Y onto X), it can be done at smaller sample sizes, with further samples collected for estimation only if an effect is detected by the hypothesis test. For such situations, correlation only tests for univariate linear independence, while other statistics like mutual information that do characterize multivariate independence are hard to estimate from data, suffering from a serious curse of dimensionality. A recent popular approach for this problem (and a related two-sample testing problem) involve the use of quantities defined in reproducing kernel Hilbert spaces (RKHSs) - see [37, 39, 40, 50].

This chapter will concern itself with increasing the statistical power at small samples of a popular kernel statistic called HSIC, by using *shrunk* empirical estimators of the unknown population quantity (introduced below).

E.1.1 Hilbert Schmidt Independence Criterion

Due to limited space, familiarity with RKHS terminology is assumed - see [110] for an introduction. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be two positive-definite reproducing kernels that correspond to RKHSs $\mathcal{H}_k$ and $\mathcal{H}_l$ respectively with inner-products $\langle \cdot, \cdot \rangle_k$ and $\langle \cdot, \cdot \rangle_l$. Let k, l arise from (implicit) feature maps $\phi : \mathcal{X} \to \mathcal{H}_k$ and $\psi : \mathcal{Y} \to \mathcal{H}_l$. In other words, ϕ, ψ are not functions, but mappings to the Hilbert space. i.e. $\phi(x) \in \mathcal{H}_k, \psi(y) \in \mathcal{H}_l$ respectively. These functions, when evaluated at points in the original spaces, must satisfy $\phi(x)(x') = \langle \phi(x), \phi(x') \rangle_k = k(x, x')$ and $\psi(y)(y') = \langle \psi(y), \psi(y') \rangle_l = l(y, y')$.

The mean embedding of P_X and P_Y are defined as $\mu_X := \mathbb{E}_{x \sim P_X} \phi(x) \in \mathcal{H}_k$ and $\mu_Y := \mathbb{E}_{y \sim P_Y} \psi(y) \in \mathcal{H}_l$ whose empirical estimates are $\widehat{\mu}_X := \frac{1}{n} \sum_{i=1}^n \phi(x_i)$ and $\widehat{\mu}_Y := \frac{1}{n} \sum_{i=1}^n \psi(y_i)$. Finally, the cross-covariance operator of X, Y is defined as

$$\Sigma_{XY} := \mathbb{E}_{(x,y) \sim P_{XY}} (\phi(x) - \mu_X) \otimes (\psi(y) - \mu_Y)$$

where $\otimes$ is an outer-product. For unfamiliar readers, if we used the linear kernel $k(x, x') = x^T x'$ and $l(y, y') = y^T y'$, then the cross-covariance operator is just the cross-covariance matrix. The plug-in empirical estimator of Σ_{XY} is

$$S_{XY} := \frac{1}{n} \sum_{i=1}^n (\phi(x_i) - \widehat{\mu}_X) \otimes (\psi(y_i) - \widehat{\mu}_Y)$$

For conciseness, define $\widetilde{\phi}(x_i) = \phi(x_i) - \widehat{\mu}_X$, $\widetilde{\psi}(y_i) = \psi(y_i) - \widehat{\mu}_Y$, $\widetilde{k}(x, x') = \langle \widetilde{\phi}(x), \widetilde{\phi}(x') \rangle_k$ and $\widetilde{l}(y, y') = \langle \widetilde{\psi}(y), \widetilde{\psi}(y') \rangle_l$. The test statistic Hilbert-Schmidt Independence Criterion (HSIC) defined in [37] is the squared Hilbert-Schmidt norm of S_{XY}, and can be calculated using centered

kernel matrices $\widetilde{K}, \widetilde{L}$, where $\widetilde{K}_{ij} = \widetilde{k}(x_i, x_j), \widetilde{L}_{ij} = \widetilde{l}(y_i, y_j)$, as

$$\text{HSIC} := \|S_{XY}\|_{HS}^2 = \frac{1}{n^2}\text{tr}(\widetilde{K}\widetilde{L}) \tag{E.1}$$

For unfamiliar readers, if we used the linear kernel, this just corresponds to the Frobenius norm of the cross-covariance matrix. The most important property is: *when the kernels k, l are "characteristic", then the corresponding population statistic $\|\Sigma_{XY}\|_{HS}^2$ is zero iff X, Y are independent* [37]. This gives rise to a natural test - calculate $\|S_{XY}\|_{HS}^2$ and reject the null if it is large.

Examples of characteristic kernels include Gaussian $k(x, x') = \exp\left(-\frac{\|x-x'\|_2^2}{\gamma^2}\right)$ and Laplace $k(x, x') = \exp\left(-\frac{\|x-x'\|_1}{\gamma}\right)$, for any bandwidth γ, while the aforementioned linear kernel is not characteristic — the corresponding HSIC tests only linear relationships, and a zero cross-covariance matrix characterizes independence only for multivariate Gaussian distributions. Working with the infinite dimensional operator with characteristic kernels, allows us to identify any general nonlinear dependence (in the limit) between any pair of distributions, not just Gaussians.

E.1.2 Independence Testing using HSIC

A permutation-based test is described in [37], and proceeds in the following manner. From the given data, calculate the test statistic $T := \|S_{XY}\|_{HS}^2$. Keeping the order of $x_1, ..., x_n$ fixed, randomly permute $y_1, ..., y_n$ a large number of times, and recompute the *permuted* HSIC each time. This destroyed any dependence between x, y simulating a draw from the product of marginals, making the empirical distribution of the permuted HSICs behave like the null distribution of the test statistic (distribution of HSIC when $\mathcal{H}_0$ is true). For a pre-specified type-1 error α, calculate threshold t_α in the right tail of the null distribution. Reject $\mathcal{H}_0$ if $T > t_\alpha$. This test was proved to be *consistent* against any fixed alternative, meaning for any fixed type-1 error α, the power goes to 1 as $n \to \infty$. Empirically, the power can be calculated using simulations by repeating the above permutation test many times for a fixed P_{XY} (for which dependence holds), and reporting the empirical probability of rejecting the null (detecting the dependence). Note that the power depends on P_{XY} (unknown to the user of the test).

E.1.3 Shrunk Estimators of S_{XY}

Even though S_{XY} is an unbiased estimator of Σ_{XY}, it typically has high variance at low sample sizes. The idea of Stein shrinkage [116] is to trade-off bias and variance, first introduced in the context of Gaussian mean estimation. This strategy of introducing some bias and decreasing the variance to get different estimators of Σ_{XY} was followed by [82] who define a linear shrinkage estimator of S_{XY} called SCOSE (Simple Covariance Shrinkage Estimator) and a nonlinear shrinkage estimator called FCOSE (Flexible Covariance Shrinkage Estimator). When we refer to shrunk estimators, we implicitly mean SCOSE and FCOSE. We will describe these briefly in Section 2.

E.1.4 Contributions

Our first contribution is the following :

1. We provide evidence that employing shrunk estimators of Σ_{XY}, instead of S_{XY}, to calculate the aforementioned test statistic, can increase the power of the associated independence test at low false positive rates, when the sample size is small (there is higher variance in estimating infinite-dimensional operators).

Our second contribution is to analyze the effect of shrinkage on the test statistic, to provide some practical insight.

2. The effect of shrinkage on the test-statistic is very similar to soft-thresholding (see Section 4), shrinking very small statistics to zero, and shrinking other values nearly (but not) linearly, and nearly (but not) monotonically.

Our last contribution is an insight on the two estimators considered in this chapter, SCOSE and FCOSE.

3. We prove that SCOSE is (essentially, up to lower order terms) the optimal/oracle linear shrinkage estimator with respect to quadratic risk (see Section 5). However, we observe that FCOSE typically achieves higher power than SCOSE. This indicates that it may be useful to search for the optimal estimator in a larger class than linearly shrunk estimators, and also that quadratic loss may not be the right loss function for the purposes of test power.

The rest of this chapter is organized as follows. Section 2 introduces SCOSE, FCOSE and their corresponding shrunk test statistics. Section 3 presents illuminating experiments that bring out the statistically significant improvement in power over HSIC. Section 4 conducts a deeper investigation into the effect of shrinkage and proves the oracle optimality of SCOSE under quadratic risk.

E.2 Shrunk Estimators and Test Statistics

Let $\mathcal{HS}(\mathcal{H}_k, \mathcal{H}_l)$ represent the set of Hilbert-Schmidt operators from $\mathcal{H}_k$ to $\mathcal{H}_l$. We first note that S_{XY} can be written as the solution to the following optimization problem.

$$S_{XY} := \min_{Z \in \mathcal{HS}(\mathcal{H}_k, \mathcal{H}_l)} \frac{1}{n} \sum_{i=1}^{n} \left\| \widetilde{\phi}(x_i) \otimes \widetilde{\psi}(y_i) - Z \right\|_{HS}^2$$

Using this idea [82] suggest the following two shrunk/regularized estimators.

From SCOSE to HSIC^S

This is derived in [82] by solving

$$\min_{Z \in \mathcal{HS}(\mathcal{H}_k, \mathcal{H}_l)} \frac{1}{n} \sum_{i=1}^{n} \left\| \widetilde{\phi}(x_i) \otimes \widetilde{\psi}(y_i) - Z \right\|_{HS}^2 + \lambda \|Z\|_{HS}^2$$

and the optimal solution (called SCOSE) is

$$S_{XY}^S := \left(1 - \frac{\lambda}{1+\lambda}\right) S_{XY}$$

where λ (and hence the shrinkage intensity) is estimated by leave-one-out cross-validation (LOOCV), in closed form as

$$\begin{aligned}\rho^S &:= \left(\frac{\lambda^{CV}}{1+\lambda^{CV}}\right) \\ &= \frac{\left[\frac{1}{n}\sum_{i=1}^n \widetilde{K}_{ii}\widetilde{L}_{ii} - \frac{1}{n^2}\sum_{i,j=1}^n \widetilde{K}_{ij}\widetilde{L}_{ij}\right]}{(n-2)\frac{1}{n^2}\sum_{i,j=1}^n \widetilde{K}_{ij}\widetilde{L}_{ij} + \frac{1}{n^2}\sum_{i=1}^n \widetilde{K}_{ii}\widetilde{L}_{ii}}\end{aligned}$$

Observing the expression for λ^{CV} in [82], the denominator can be negative (for example, with the Gaussian kernel for small bandwidths, resulting in a kernel matrix close to the identity). This can cause λ^{CV} to be negative, and ρ^S to be (unintentionally) outside the range $[0,1]$. Though not discussed in [82], we shall follow the convention that when $\rho^S < 0$, we shall use $\rho^S = 0$ and if $\rho^S > 1$, we use $\rho_S = 1$. Indeed, one can show that $\left(1 - \frac{\lambda}{1+\lambda}\right)_+ S_{XY}$ dominates $\left(1 - \frac{\lambda}{1+\lambda}\right) S_{XY}$ where $(x)_+ = \max\{x, 0\}$. In Section 4, we prove that S^S_{XY} is (essentially) the optimal/oracle linear shrinkage estimator with respect to quadratic risk.

We can now calculate the corresponding shrunk statistic HSIC^S $= \|S^S_{XY}\|^2_{HS} =$

$$\left(1 - \frac{\frac{1}{n}\sum_{i=1}^n \widetilde{K}_{ii}\widetilde{L}_{ii} - \text{HSIC}}{(n-2)\text{HSIC} + \frac{\frac{1}{n}\sum_{i=1}^n \widetilde{K}_{ii}\widetilde{L}_{ii}}{n}}\right)_+^2 \text{HSIC} \tag{E.2}$$

While the above expression looks daunting, one thing to note is that the amount that HSIC is shrunk (i.e. the multiplicative factor) depends on the value of HSIC. As we shall see in section 4, small HSIC values get shrunk to zero, but as can be seen above, the shrinkage of HSIC is non-monotonic.

From FCOSE to HSIC^F

The Flexible Covariance Shrinkage Estimator is derived by relying on the Representer theorem, see [110], to instead minimize

$$\frac{1}{n}\sum_{i=1}^n \left\| \widetilde{\phi}(x_i) \otimes \widetilde{\psi}(y_i) - \sum_{i=1}^n \frac{\beta_i}{n}\widetilde{\phi}(x_i) \otimes \widetilde{\psi}(y_i) \right\|^2_{HS} + \lambda\|\beta\|_2^2$$

over all $\beta \in \mathbb{R}^n$, and the optimal solution (called FCOSE) is

$$\begin{aligned} S^F_{XY} &:= \sum_{i=1}^n \frac{\beta_i^\lambda}{n}\widetilde{\phi}(x_i) \otimes \widetilde{\psi}(y_i) \\ \text{where } \beta^\lambda &= (\widetilde{K} \circ \widetilde{L} + \lambda I)^{-1}\widetilde{K} \circ \widetilde{L}\mathbf{1} \end{aligned}$$

where $\circ$ denotes elementwise (Hadamard) product, $\mathbf{1}$ is the vector $[1, 1, ..., 1]^T$, and as before the best λ is determined by LOOCV. The procedure to evaluate the optimal λ efficiently is described by [82] - a single eigenvalue decomposition of $\widetilde{K} \circ \widetilde{L}$ costing $O(n^3)$ can be done, following which

evaluating LOOCV is only $O(n^2)$ per λ, see [82], section 3.1 for more details. As before, after picking the λ by LOOCV, we can derive the corresponding shrunk test statistic as

$$\begin{aligned}\text{HSIC}^F &= \|S^S_{XY}\|^2_{HS} \\ &= \frac{1}{n^2}\text{tr}(M(M+\lambda I)^{-1}M(M+\lambda I)^{-1}M)\end{aligned}$$

where $M = \widetilde{K} \circ \widetilde{L}$. Note here that the shrinkage is not linear, and the effect on HSIC cannot be seen immediately. Similar to SCOSE, we shall see in section 4, small HSIC values get shrunk to zero (LOOCV chooses a large λ).

E.3 Linear Shrinkage and Quadratic Risk

In this section, we prove that SCOSE is (essentially) optimal within a particular class of estimators. Such "oracle" arguments also exist elsewhere in the literature, like [69], so we provide only a brief proof outline.

Proposition 1. *The oracle (with respect to quadratic risk) linear shrinkage estimator and intensity is defined as*

$$S^*, \rho^* := \underset{Z\in\mathcal{HS}, Z=(1-\rho)S_{XY}, 0\le\rho\le 1}{\text{argmin}} \|Z - \Sigma_{XY}\|^2_{HS}$$

and is given by $S^* := (1-\rho^*)S_{XY}$ *where*

$$\rho^* := \frac{\mathbb{E}\|S_{XY} - \Sigma_{XY}\|^2_{HS}}{\mathbb{E}\|S_{XY}\|^2}$$

Proof. Define $\alpha^2 = \|\Sigma_{XY}\|^2_{HS}$, $\beta^2 = \mathbb{E}\|S_{XY} - \Sigma_{XY}\|^2_{HS}$, $\delta^2 = \mathbb{E}\|S_{XY}\|^2$. Since $\mathbb{E}[S_{XY}] = \Sigma_{XY}$, it is easy to verify that $\alpha^2 + \beta^2 = \delta^2$. Substituting and expanding the objective, we get:

$$\begin{aligned}\mathbb{E}\|Z - \Sigma_{XY}\|^2_{HS} &= \mathbb{E}\| - \rho S_{XY} + (S_{XY} - \Sigma_{XY})\|^2_{HS} \\ &= \rho^2\delta^2 + \beta^2 - 2\rho(\delta^2 - \alpha^2) \\ &= \rho^2\alpha^2 + (1-\rho)^2\beta^2\end{aligned}$$

Differentiating and equating to zero, gives $\rho^* = \frac{\beta^2}{\delta^2}$. □

This ρ^* appears in terms of quantities that depend on the unknown underlying distribution (hence the term *oracle* estimator). We use plugin estimates b, d for β, δ.

Let $d^2 = \|S_{XY}\|^2_{HS} = \frac{1}{n^2}\sum_{i,j=1}^n \widetilde{K}_{ij}\widetilde{L}_{ij} = HSIC$. Since β^2 is the variance of S_{XY}, let b^2 be the sample variance of S_{XY}, i.e.

$b^2 = \frac{1}{n}\frac{1}{n}\sum_{k=1}^n ||\widetilde{\phi}(x_i)\otimes\widetilde{\psi}_{x_i} - S_{XY}||^2 = \frac{1}{n}\left[\frac{1}{n}\sum_{i=1}^n \widetilde{K}_{ii}\widetilde{L}_{ii} - \frac{1}{n^2}\sum_{i,j=1}^n \widetilde{K}_{ij}\widetilde{L}_{ij}\right]$.

Plugging these into S^* and simplifying, we see that $\text{HSIC}^* := \|S^*\|^2_{HS}$ is

$$\text{HSIC}^* = \left(1 - \frac{\frac{1}{n}\sum_{i=1}^n \widetilde{K}_{ii}\widetilde{L}_{ii} - \text{HSIC}}{n\text{HSIC}}\right)^2 \text{HSIC} \tag{E.3}$$

Comparing Eq.(E.3) with Eq.(E.2) shows that SCOSE is essentially S^*, up to a factor in the denominator which is of the same order as the bias of the HSIC empirical estimator[1] (see Theorem 1 in [37]). In other words, SCOSE just corresponds to using a slightly different estimator for δ^2 than the simple plugin d^2, which varies on the same order as the bias $\delta^2 - \mathbb{E}d^2$. Hence SCOSE, as estimated via regularization and LOOCV, is (essentially) the optimal linear shrinkage estimator under quadratic risk.

To the best of our knowledge, this is the first such characterization of *optimality of an estimator achieved through leave-one-out cross-validation*. We are only able to prove this because one can explicitly calculate both the oracle linear shrinkage intensity ρ^* as well as the optimal λ^{CV} (as mentioned in Section 2). This raises a natural open question — can we find other situations where the LOOCV estimator is optimal with respect to some risk measure? (perhaps when explicit calculations are not possible, like ridge regression).

E.4 Experiments

In this section, we run three kinds of experiments: a) to verify that SCOSE has better quadratic risk than FCOSE and original sample estimator, b) detailed synthetic experiments to verify that shrinkage does improve power, across interesting regimes of $\alpha = \{0.01, 0.05, 0.1\}$, and c) real data obtained from MNIST, to show that we shrinkage detect dependence at much lower samples than the original data size.

E.4.1 Quadratic Risk

Figure E.1 shows that SCOSE is indeed much better than both S_{XY} and FCOSE with respect to quadratic risk. Here, we calculate $\mathbb{E}\|Z - \Sigma_{XY}\|_{HS}^2$ for the distribution given in dataset (A) for $Z \in \{S_{XY}, S_{XY}^S, S_{XY}^F\}$. The expectation is calculated by repeating the experiment 1000 times. Each time Z is calculated according to $N \in \{20, 50, 100\}$ samples and Σ_{XY} is approximated by the empirical cross-covariance matrix on 5,000 samples. The four panels use four different kernels which are linear, polynomial, Laplace and Gaussian from top to bottom. The shrunk estimators are always better than the unshrunk, with a larger difference between SCOSE and FCOSE for finite-dimensional feature spaces (top two). In infinite-dimensional feature spaces (bottom two), SCOSE and FCOSE are much better than the unshrunk estimator but very similar to each other. The differences between all estimators decreases with increasing n, since the sample cross-covariance operator itself becomes very accurate.

E.4.2 Synthetic Data

We perform synthetic experiments in a wide variety of settings to demonstrate that the shrunk test statistics achieve higher power than HSIC in a variety of settings. We follow the schema provided in the introduction for independence testing and calculating power. We only consider difficult distributions with nonlinear dependence between X, Y, on which linear methods like

[1] HSIC and HSIC$-$2HSIC$/n - C/n^2$ both converge to population HSIC at same rate determined by the dominant term (HSIC).

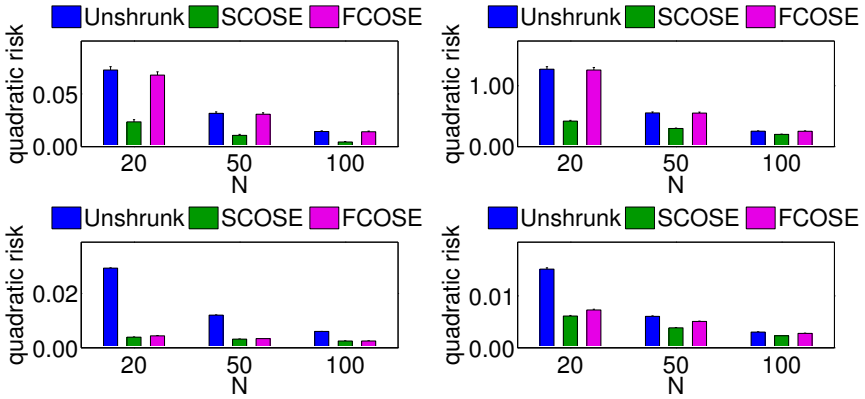


Figure E.1: All panels show quadratic risk $\mathbb{E}\|X - \Sigma_{XY}\|^2_{HS}$ for $X \in \{S_{XY}, S^S_{XY}, S^F_{XY}\}$. Dataset (A) was used in all four panels, but the kernels were varied - from top to bottom is the linear, quadratic, Gaussian and Laplace kernel.

correlation are shown to fail to detect dependence (some of them were used in previous papers on independence testing like [42] and [14]).

For all experiments, $\alpha \in \{0.01, 0.05, 0.1\}$ is chosen as the type-1 error (for choosing the threshold level of the null distribution's right tail). For every setting of parameters of each experiment, power is calculated as the percentage of rejection over 200 repetitions (independent trials), with 2000 permutations per repetition (permutation testing to find the null distribution threshold at level α). We use the Gaussian kernel where the bandwidth is chosen by the common median heuristic [110].

Table E.1 is a representative sample from what we saw on other examples - either large, small or no improvement in power but almost never a decrease. The improvements may not always be huge, but they are statistically significant - it is difficult to detect such non-linear dependencies at low sample sizes, so *any* increase in power can be important in scientific applications.

Remark. A more appropriate way than using error bars to assess significance is by the Wilcoxon rank sum test, omitted for lack of space, though it yields more favorable results.

E.4.3 Real Data

We use two real datasets - the first is a good example where shrinkage helps a lot, but in the second it does not help (we show it on purpose). Like the synthetic datasets, for most real datasets it either helps or does not hurt (being very rarely worse; see remark in the discussion).

The first is the Eckerle dataset [20] from the NIST Statistical Reference Datasets (NIST StRD) for Nonlinear Regression, data from a NIST study of circular interference transmittance (n=35, Y is transmittance, X is wavelength). A plot of the data in Figure E.2 reveals a nonlinear relationship between X, Y (though the correlation is 0.035 with p-value 0.84). We subsample the data to see how often we can detect a relationship at $10\%, 20\%, 30\%$ of the original data size, when the false positive level is always controlled at 0.05. The second is the Aircraft dataset [8] (n=709, X is log(speed), Y is log(span)). Once again, correlation is low, with a p-value of over 0.8, and we subsample the data to $5\%, 10\%, 20\%$ of the original data size.

	$\alpha = 0.01$					$\alpha = 0.05$					$\alpha = 0.10$				
	HSIC	HSIC$_S$	HSIC$_F$			HSIC	HSIC$_S$	HSIC$_F$			HSIC	HSIC$_S$	HSIC$_F$		
	0.22 ±0.03	0.21 ±0.03	0.34 ±0.03		✓	0.52 ±0.04	0.52 ±0.04	0.71 ±0.03		✓	0.73 ±0.03	0.72 ±0.03	0.90 ±0.02		✓
	0.41 ±0.03	0.41 ±0.03	0.48 ±0.04		✓	0.68 ±0.03	0.68 ±0.03	0.88 ±0.02		✓	0.85 ±0.03	0.85 ±0.02	0.99 ±0.01		✓
	0.41 ±0.03	0.40 ±0.03	0.52 ±0.04		✓	0.74 ±0.03	0.74 ±0.03	0.94 ±0.02		✓	0.94 ±0.02	0.94 ±0.02	0.99 ±0.01		✓
	0.52 ±0.04	0.52 ±0.04	0.66 ±0.03		✓	0.91 ±0.02	0.91 ±0.02	0.89 ±0.02			0.99 ±0.01	0.99 ±0.01	0.96 ±0.01		✗
	0.04 ±0.01	0.04 ±0.01	0.04 ±0.01			0.12 ±0.02	0.12 ±0.02	0.14 ±0.02			0.23 ±0.03	0.23 ±0.03	0.24 ±0.03		
	0.10 ±0.02	0.10 ±0.02	0.12 ±0.02			0.31 ±0.03	0.31 ±0.03	0.40 ±0.03		✓	0.47 ±0.04	0.47 ±0.04	0.58 ±0.03		✓
	0.33 ±0.03	0.33 ±0.03	0.46 ±0.04		✓	0.77 ±0.03	0.77 ±0.03	0.91 ±0.02		✓	0.95 ±0.01	0.96 ±0.01	0.99 ±0.01		✓
	0.93 ±0.02	0.93 ±0.02	0.96 ±0.01		✓	1.00 ±0.00	1.00 ±0.00	1.00 ±0.00			1.00 ±0.00	1.00 ±0.00	1.00 ±0.00		
	0.07 ±0.02	0.07 ±0.02	0.09 ±0.02			0.24 ±0.03	0.26 ±0.03	0.32 ±0.03		✓	0.44 ±0.04	0.47 ±0.04	0.48 ±0.04		
	0.06 ±0.02	0.07 ±0.02	0.09 ±0.02			0.26 ±0.03	0.28 ±0.03	0.32 ±0.03			0.45 ±0.04	0.47 ±0.04	0.48 ±0.04		
	0.10 ±0.02	0.12 ±0.02	0.14 ±0.02			0.34 ±0.03	0.34 ±0.03	0.39 ±0.03			0.51 ±0.04	0.52 ±0.04	0.53 ±0.04		
	0.07 ±0.02	0.07 ±0.02	0.10 ±0.02		✓	0.30 ±0.03	0.33 ±0.03	0.35 ±0.03			0.53 ±0.04	0.54 ±0.04	0.57 ±0.04		
	0.04 ±0.01	0.05 ±0.02	0.04 ±0.01			0.18 ±0.03	0.27 ±0.03	0.24 ±0.03	✓	✓	0.34 ±0.03	0.45 ±0.04	0.44 ±0.04	✓	✓
	0.16 ±0.03	0.20 ±0.03	0.20 ±0.03			0.45 ±0.04	0.58 ±0.03	0.58 ±0.03	✓	✓	0.67 ±0.03	0.73 ±0.03	0.73 ±0.03	✓	✓
	0.34 ±0.03	0.43 ±0.04	0.43 ±0.04	✓	✓	0.71 ±0.03	0.80 ±0.03	0.79 ±0.03	✓	✓	0.85 ±0.03	0.90 ±0.02	0.89 ±0.02	✓	
	0.63 ±0.03	0.72 ±0.03	0.73 ±0.03	✓	✓	0.91 ±0.02	0.92 ±0.02	0.92 ±0.02			0.95 ±0.01	0.96 ±0.01	0.96 ±0.01		

Table E.1: The first column shows scatterplots of X vs Y (all having dependence between X, Y). There are 3 sets of 5 columns each - for $\alpha = 0.01, 0.05, 0.1$ (controlled by running 2000 permutations). In eachs set, the first three columns show the power of HSIC, HSIC^S, HSIC^F (with standard deviation over 200 repetitions below). The fourth column shows when HSIC^S is significantly better than HSIC, and the fifth column when HSIC^F has significantly higher power than HSIC. A blank means the powers are not significantly better or worse. In the first dataset (A) (top 4) we show how the power varies with increasing n (becomes easier). In the second dataset (B) (second 4) we show how the power varies with rotation (goes from near-independence to clear dependence). In the third dataset (C) (third 4), we demonstrate a case where shrinkage does *not* help much, which is a circle with a hole. In the last dataset (D) (last 4), we demonstrate a case where HSIC^S does as well as HSIC^F. We tried many more datasets, these are a few representative samples.

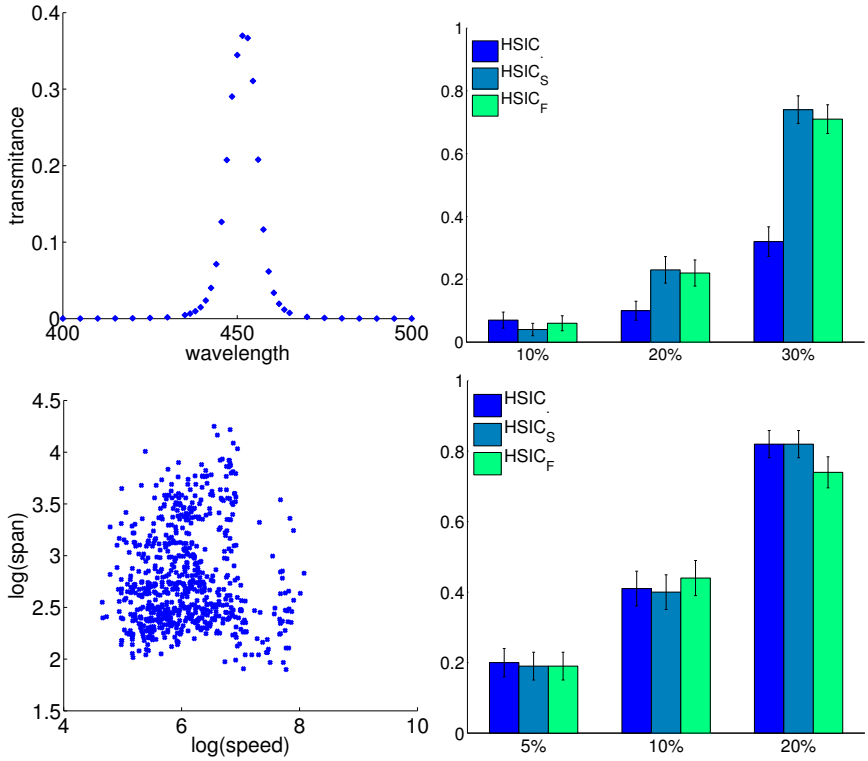


Figure E.2: Top Row: The left figure shows a plot of wavelength against transmittance. The right figure shows the power of HSIC, HSIC^S, HSIC^F when the data are subsampled to $10\%, 20\%, 30\%$ (error bars over 100 repetitions). Bottom Row: The left figure shows a plot of $log(wingspan)$ vs $log(airspeed)$. The right figure shows the power of HSIC, HSIC^S, HSIC^F when the data are subsampled to $5\%, 10\%, 20\%$ (error bars over 100 repetitions).

E.5 Discussion

Why might shrinkage improve power? Let us examine the net effect of using shrunk estimators on the value of HSIC, i.e. let us compare HSIC^S and HSIC^F to HSIC by computing these over all the repetitions of the permutation testing procedure described in the introduction. In Fig. E.3, both estimators are visually similar in transforming the actual test statistic. Perhaps the more interesting phenomenon is that Fig. E.3 is reminiscent of the graph of a soft-thresholding operator $ST_t(x) = \max\{0, x-t\}$. Intuitively, if the unshrunk HSIC value is small, the shrinkage methods deem it to be "noise" and it is shrunk to zero. Looking at the X-axis scaling of the top and bottom row, the size of the region that gets shrunk to zero decreases with n - as expected, shrinkage has less effect when S_{XY} has low variance). The shrinkage being non-monotone (more so for $n = 20$ than $n = 50$ in Figure E.3) is key to achieving an improvement in power.

Using the intuition from the above figure, we can finally piece together why shrinkage may yield benefits. A rejection of $\mathcal{H}_0$ occurs when the test statistic stands out in the right tail of its null distribution. Typically, when the alternative is true (this is when rejecting the null improves power) the unshrunk test statistics calculated from the permuted samples is smaller than the

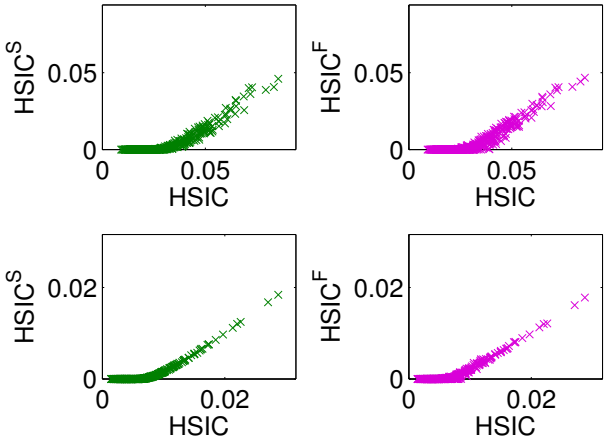


Figure E.3: The top row corresponds to $n = 20$, and the bottom row has $n = 50$. The left plots compare HSIC^S to HSIC, and the right plots compare HSIC^F to HSIC. Each cross mark corresponds to the shrunk and unshrunk HSIC calculated during a single permutation of a permutation test.

unshrunk HSIC calculated on the original sample. However, the effect of shrinking the small statistics towards zero, and setting the smallest ones to zero, is that the unpermuted test statistic under the alternative distribution stands out more in the right tail of the null.

In other words, relative to the unshrunk null distribution and the unshrunk test statistic, the tail of the null distribution is shrunk more towards zero than the unpermuted test statistic, causing the latter to have a higher quantile in the right tail of the former (relative to the quantile before shrinkage). Let us verify this experimentally. In Fig.E.4 we plot for each of the datasets in Table E.1, the average ratio of unpermuted statistic T to the 95th percentile of the permuted statistics, for $T \in \{\text{HSIC}, \text{HSIC}^S, \text{HSIC}^F\}$. Recall that for dataset (C), we didn't see much of an improvement in power, but for (A),(B),(D) it is clear from Fig. E.4 that the unpermuted statistic is shrunk less than its null distribution's 95th quantile.

Remark. In our experiments, real and synthetic, shrinkage usually improves (and almost never worsens) power in false-positive regimes that we usually care about. Will shrinkage *always* improve power? Possibly not. Even though shrunk the shrunk S_{XY} dominates S_{XY} for estimation error, it may not be the case that shrunk HSIC always dominates unshrunk HSIC for test power (i.e. the latter may not be *inadmissible*). However, just as no single classifier always outperforms another, it is still beneficial to add techniques like shrinkage, that seem to consistently yield benefits in practice, to the practitioner's array of tools.

E.6 Conclusion

We presented evidence for an important phenomenon - using biased but lower variance shrunk estimators of cross-covariance operators can often significantly improve test power of HSIC at small sample sizes. This observation (that shrinkage can improve power) has rarely been made in the statistics and machine learning testing literature. We think the reason is that most test statistics for independence testing cannot be immediately expressed as the norm of an empirical

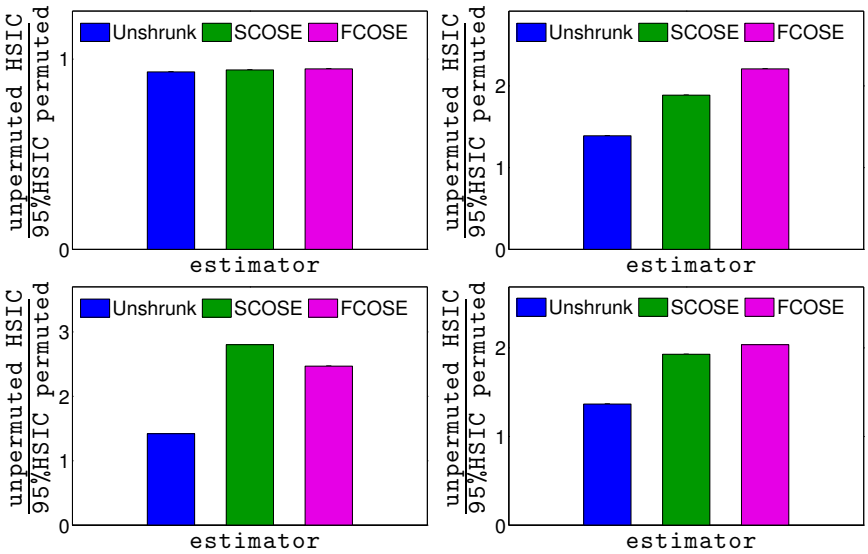


Figure E.4: All panels show the ratio of the unpermuted HSIC to the 95th percentile of the null distribution based on HSICs calculated from the permuted data. (see Table E.1) The top row has datasets (C) with radius 2.2, (B) with angle $3 \times \pi/32$, and the bottom row has (D) with $N = 25$, (A) with $N = 40$. These observations were qualitatively the same in all other synthetic data parameter settings, and also for other percentiles than 95th, and since the figures look identical in spirit, they were omitted due to lack of space.

operator, making it less obvious *how* to apply shrinkage to improve their power at low sample sizes.

We also showed the optimality (among linear shrinkage estimators) of SCOSE, but observe that the nonlinear shrinkage of FCOSE usually yields higher power. To the best of our knowledge, there seems to be no current literature showing that the choice made by leave-one-out cross-validation (SCOSE) explicitly leads to an estimator that is "optimal" in some sense (among linear shrinkage estimators). This may be because it is often not possible to explicitly calculate the form of the LOOCV estimator, nor the explicit form of the best linear shrinkage estimator, as can both be done in this simple setting.

Since even the best possible linear shrinkage estimator (as represented by SCOSE) is usually worse than FCOSE, this result indicates that in order to improve upon FCOSE, it will be necessary to further study the class of non-linear shrinkage estimators for our infinite dimensional operators, as done for finite dimensional covariance matrices in [70] and other papers by the same authors.

We ended with a brief investigation into the effect of shrinkage on HSIC and why shrinkage may intuitively improve power. We think that our work will be important for more powerful non-parametric detection of subtle nonlinear dependencies at low sample sizes, a common problem in scientific applications.

Bibliography

[1] Adrian Akmajian. *Linguistics: An introduction to language and communication*. MIT press, 2001. 4.2.1, 4.2.1, 4.2.1

[2] John R Anderson, Hee Seung Lee, and Jon M Fincham. Discovering the structure of mathematical problem solving. *NeuroImage*, 97:163–177, 2014. 1, 2.3

[3] J. Ashburner, CC Chen, G. Flandin, R. Henson, S. Kiebel, J. Kilner, V. Litvak, R. Moran, W. Penny, K. Stephan, et al. SPM8 manual. *Functional Imaging Laboratory, Institute of Neurology*, 2008. 3.5.3, 4.1

[4] F. Gregory Ashby. *Statistical Analysis of fMRI Data*. MIT Press, Cambridge, Massachusetts, 2011. 1, 3.5, 3.5.3

[5] Asaf Bachrach. *Imaging neural correlates of syntactic complexity in a naturalistic context*. PhD thesis, Massachusetts Institute of Technology, 2008. 2.3.2

[6] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Annals of statistics*, pages 1165–1188, 2001. 3.4.2, ?

[7] Jelmer P Borst and John R Anderson. The discovery of processing stages: Analyzing eeg data with hidden semi-markov models. *NeuroImage*, 108:60–73, 2015. 7.1.1

[8] A. W. Bowman and A. Azzalini. *R package `sm`: nonparametric smoothing methods (version 2.2-5.4)*. University of Glasgow, UK and Università di Padova, Italia, 2014. URL `URLhttp://www.stats.gla.ac.uk/~adrian/sm,http://azzalini.stat.unipd.it/Book_sm`. E.4.3

[9] D.H. Brainard. The psychophysics toolbox. *Spatial vision*, 10(4):433–436, 1997. 4.1

[10] J. Brennan, Y. Nir, U. Hasson, R. Malach, D.J. Heeger, and L. Pylkkänen. Syntactic structure building in the anterior temporal lobe during natural story listening. *Brain and language*, 2010. 2.3, 2.3.2, 4.2.2

[11] Tamara Broderick, Nicholas Boyd, Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. Streaming variational Bayes. In C. J. C. Burges, Léon Bottou, Max Welling, Zoubin Ghahramani, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26 [NIPS 2013]*, pages 1727–1735, 2013. URL `http://papers.nips.cc/paper/4980-streaming-variational-bayes`. 3.5.5

[12] A. Buchweitz, R.A. Mason, L. Tomitch, and M.A. Just. Brain activation for reading and listening comprehension: An fMRI study of modality effects and individual differences in language comprehension. *Psychology & Neuroscience*, 2(2):111–123, 2009. 4.1,

5.3.1

[13] Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. *arXiv preprint arXiv:1402.4501*, 2014. 3.4.2

[14] Kacper Chwialkowski and Arthur Gretton. A kernel independence test for random processes. In *Proceedings of The 31st International Conference on Machine Learning*, pages 1422–1430, 2014. 6.3, E.4.2

[15] Laurent Cohen and Stanislas Dehaene. Specialization within the ventral stream: the case for the visual word form area. *Neuroimage*, 22(1):466–476, 2004. 4.2.2, 4.4

[16] R.T. Constable, K.R. Pugh, E. Berroya, W.E. Mencl, M. Westerveld, W. Ni, and D. Shankweiler. Sentence complexity and input modality effects in sentence comprehension: an fMRI study. *Neuroimage*, 22(1):11–21, 2004. 1

[17] G. Currie. *The Nature of Fiction*. Cambridge University Press, 1990. 1

[18] M. Dapretto and S.Y. Bookheimer. Form and content: dissociating syntax and semantics in sentence comprehension. *Neuron*, 24(2):427–432, 1999. 2.1.2

[19] Marie-Catherine De Marneffe and Christopher D Manning. Stanford typed dependencies manual. Technical report, Technical report, Stanford University, 2008. 4.2.1

[20] K Eckerle. Circular interference transmittance study. *National Institute of Standards and Technology (NIST), US Department of Commerce, USA*, 1979. E.4.3

[21] E. Fedorenko, P.-J. Hsieh, A. Nieto-Castanon, S. Whitfield-Gabrieli, and N. Kanwisher. New method for fMRI investigations of language: Defining ROIs functionally in individual subjects. *Journal of Neurophysiology*, 104(2):1177–1194, 2010. 1, D

[22] E. Fedorenko, A. Nieto-Castanon, and N. Kanwisher. Lexical and syntactic representations in the brain: An fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia*, 50(4):499–513, 2012. 2.1.2, 4.4, 4.4, 4.4, 5.7

[23] Evelina Fedorenko. The role of domain-general cognitive control in language comprehension. *Frontiers in psychology*, 5, 2014. 4.4

[24] Evelina Fedorenko and Sharon L Thompson-Schill. Reworking the language network. *Trends in cognitive sciences*, 18(3):120–126, 2014. 2.1.2, 2.6, 4.4

[25] Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. Word surprisal predicts N400 amplitude during reading. In *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*, pages 878–883, 2013. 2.1.1, 5.1, 3

[26] Angela D Friederici. Towards a neural basis of auditory sentence processing. *Trends in cognitive sciences*, 6(2):78–84, 2002. 1, 5.1

[27] Angela D Friederici. The brain basis of language processing: from structure to function. *Physiological reviews*, 91(4):1357–1392, 2011. 7.1.2

[28] Angela D Friederici. The cortical language circuit: from auditory perception to sentence comprehension. *Trends in cognitive sciences*, 16(5):262–268, 2012. 1, 2.4

[29] Jerome Friedman, Trevor Hastie, Robert Tibshirani, and H. Jiang. *Glmnet for Matlab*. Statistics Department, Stanford University, 2010. URL `http://www.stanford.`

edu/~hastie/glmnet_matlab/. 3.5.1, 3.5.4

[30] Karl J. Friston, Pia Rothshtein, Joy J. Geng, Philipp Sterzer, and Rik N. Henson. A critique of functional localizers. pages 3–24. 3

[31] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. Kernel measures of conditional dependence. In *NIPS*, volume 20, pages 489–496, 2007. 6.5

[32] Alona Fyshe, Partha P Talukdar, Brian Murphy, and Tom M Mitchell. Interpretable semantic vectors from a joint model of brain-and text-based meaning. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, volume 1, pages 489–499, 2014. 2.2.2, 3

[33] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. CRC Press, London, second edition, 2003. A.2

[34] M Ida Gobbini and James V Haxby. Neural systems for recognition of familiar faces. *Neuropsychologia*, 45(1):32–41, 2007. 4.4

[35] Robert F Goldberg, Charles A Perfetti, and Walter Schneider. Perceptual knowledge retrieval activates sensory brain regions. *The Journal of Neuroscience*, 26(18):4917–4921, 2006. 2.1.1

[36] Gene H Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979. 3.5.4, 2, 2c

[37] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic Learning Theory*, pages 63–77. Springer, 2005. E.1, E.1.1, E.1.1, E.1.2, E.3

[38] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with Hilbert-Schmidt norms. In *Algorithmic learning theory*, pages 63–77. Springer, 2005. 6.2

[39] Arthur Gretton, Ralf Herbrich, Alexander Smola, Olivier Bousquet, and Bernhard Schölkopf. Kernel methods for measuring independence. *The Journal of Machine Learning Research*, 6:2075–2129, 2005. E.1

[40] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Neural Information Processing Systems*, pages 513–520, 2006. E.1

[41] Arthur Gretton, Karsten M Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2006. 6.1

[42] Arthur Gretton, Kenji Fukumizu, Choon Hui Teo, Le Song, Bernhard Schölkopf, and Alex J Smola. A kernel statistical test of independence. *Neural Information Processing Systems*, 2007. E.4.2

[43] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012. 6.1, 6.1

[44] ED Grossman and R Blake. Brain activity evoked by inverted and imagined biological motion. *Vision research*, 41(10):1475–1482, 2001. 4.4

[45] P. Hagoort, G.è Baggio, and Roel M Willems. Semantic unification. *The cognitive neurosciences*, 4:819–836, 2009. 4.4

[46] Peter Hagoort. How the brain solves the binding problem for language: a neurocomputational model of syntactic processing. *Neuroimage*, 20:S18–S29, 2003. 1, 5.1

[47] Peter Hagoort. Muc (memory, unification, control) and beyond. *Frontiers in psychology*, 4, 2013. 1, 2.1, 2.3, 2.1.2

[48] Matti S Hämäläinen and RJ Ilmoniemi. Interpreting magnetic fields of the brain: minimum norm estimates. *Medical & biological engineering & computing*, 32(1):35–42, 1994. 3.3

[49] M. Hanke, Y.O. Halchenko, P.B. Sederberg, S.J. Hanson, J.V. Haxby, and S. Pollmann. Pymvpa: A Python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, 7(1):37–53, 2009. 4.1

[50] Zaïd Harchaoui, Francis Bach, and Eric Moulines. Testing for homogeneity with kernel fisher discriminant analysis. *Arxiv preprint 0804.1026*, 2007. E.1

[51] Uri Hasson and Christopher J Honey. Future trends in Neuroimaging: Neural processes as expressed within real-life contexts. *NeuroImage*, 62(2):1272–1278, 2012. 1, 2.3

[52] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, 2001. 3.5.1

[53] Stefan Haufe, Frank Meinecke, Kai Görgen, Sven Dähne, John-Dylan Haynes, Benjamin Blankertz, and Felix Bießmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87:96–110, 2014. 3.5.4, 5

[54] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, 2011. 2.3.1

[55] Gregory Hickok and David Poeppel. The cortical organization of speech processing. *Nature Reviews Neuroscience*, 8(5):393–402, 2007. 1, 2.1.2, 2.5

[56] Steven A Hillyard and Marta Kutas. Event-related potentials and magnetic fields in the human brain. *Neuropsychopharmacology: The Fifth Generation of Progress. Philadelphia, Pa: Lippincott Williams and Wilkins*, pages 427–439, 2002. 2.1.1

[57] Arthur E Hoerl and Robert W Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970. 3.5.1

[58] Andrew P Holmes, RC Blair, G Watson, and I Ford. Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism*, 16(1):7–22, 1996. 5.3.2

[59] Christopher J Honey, Thomas Thesen, Tobias H Donner, Lauren J Silbert, Chad E Carlson, Orrin Devinsky, Werner K Doyle, Nava Rubin, David J Heeger, and Uri Hasson. Slow cortical dynamics and the accumulation of information over long timescales. *Neuron*, 76

(2):423–434, 2012. 4.4

[60] R.A. Hutchinson, R.S. Niculescu, T.A. Keller, I. Rustandi, T.M. Mitchell, et al. Modeling fMRI data generated by overlapping cognitive processes with unknown onsets using Hidden Process Models. *NeuroImage*, 46(1):87–104, 2009. 3.3, 7.1.1

[61] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, 2012. 2.3.1, 4, 7.1.1

[62] M. Kleiner, D. Brainard, D. Pelli, A. Ingling, R. Murray, and C. Broussard. What's new in Psychtoolbox-3. *Perception*, 36(14):1–1, 2007. 4.1

[63] Stefan Koelsch, Thomas Fritz, Karsten Müller, Angela D Friederici, et al. Investigating emotion with music: an fmri study. *Human brain mapping*, 27(3):239–250, 2006. 7.1.1

[64] Nikolaus Kriegeskorte and Rogier A Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends in cognitive sciences*, 17(8):401–412, 2013. 6.2

[65] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103(10):3863–3868, 2006. 3.4.2, 4.3, 5.3.2, 6.4

[66] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis–connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2, 2008. 6.2

[67] G.R. Kuperberg. Neural mechanisms of language comprehension: Challenges to syntax. *Brain Research*, 1146:23–49, 2007. 1, 2.1.1

[68] Minjung Kyung, Jeff Gill, Malay Ghosh, and George Casella. Penalized regression, standard errors, and bayesian lassos. *Bayesian Analysis*, 5:369–411, 2010. 3.5.1

[69] Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. E.3

[70] Olivier Ledoit and Michael Wolf. Nonlinear shrinkage estimation of large-dimensional covariance matrices. *Institute for Empirical Research in Economics University of Zurich Working Paper*, (515), 2011. E.6

[71] H. Liu, L. Wang, and T. Zhao. Multivariate regression with calibration. In *Advances in Neural Information Processing Systems*, 2014. 6, 6.1

[72] C. D Manning and H. Schütze. *Foundations of statistical natural language processing*, volume 999. MIT Press, 1999. 1

[73] Nira Mashal, Miriam Faust, Talma Hendler, and Mark Jung-Beeman. An fmri investigation of the neural correlates underlying the processing of novel metaphoric expressions. *Brain and language*, 100(2):115–126, 2007. 4.2.1

[74] R.A. Mason and M.A. Just. Neuroimaging contributions to the understanding of discourse processes. *Handbook of psycholinguistics*, 799, 2006. 1, 2.3.2, 4.4, 4.4

[75] Ann Meyler, Timothy A Keller, Vladimir L Cherkassky, Donghoon Lee, Fumiko Hoeft,

Susan Whitfield-Gabrieli, John DE Gabrieli, and Marcel Adam Just. Brain activation during sentence comprehension among good and poor readers. *Cerebral Cortex*, 17(12): 2780–2787, 2007. 4.4

[76] Tomas Mikolov. *Statistical Language Models Based on Neural Networks*. PhD thesis, Brno University of Technology, 2012. 5.2, 5.2, 5.2

[77] Tomas Mikolov, Stefan Kombrink, Anoop Deoras, Lukar Burget, and J Cernocky. RNNLM-recurrent neural network language modeling toolkit. In *Proc. of the 2011 ASRU Workshop*, pages 196–201, 2011. 5, 5.2

[78] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995. 4.2.2

[79] T.M. Mitchell, S.V. Shinkareva, A. Carlson, K.M. Chang, V.L. Malave, R.A. Mason, and M.A. Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195, 2008. 2.1.1, 2.7, 2.2.1, 2.2.2, 3.1, 3.2, 3.3, 3.5, 3.5.1, 4, A.5

[80] Tom M Mitchell, Rebecca Hutchinson, Radu S Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. Learning to decode cognitive states from brain images. *Machine Learning*, 57(1-2):145–175, 2004. 2.2.1, 6.1

[81] Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. Predicting human brain activity associated with the meanings of nouns. *Science*, 320:1191–1195, 2008. 6.2

[82] Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, Arthur Gretton, and Bernhard Schoelkopf. Kernel mean estimation and stein effect. In *Proceedings of The 31st International Conference on Machine Learning*, pages 10–18, 2014. E.1.3, E.2, E.2, E.2

[83] B. Murphy, P. Talukdar, and T. Mitchell. Learning effective and interpretable semantic models using Non-Negative Sparse Embedding. In *International Conference on Computational Linguistics (COLING 2012), Mumbai, India*, 2012. 2.2.2, 4.2.2

[84] Brian Murphy, Massimo Poesio, Francesca Bovolo, Lorenzo Bruzzone, Michele Dalponte, and Heba Lakany. EEG decoding of semantic category reveals distributed representations for single concepts. *Brain and language*, 117(1):12–22, 2011. 4.4

[85] Brian Murphy, Partha Talukdar, and Tom Mitchell. Selecting corpus-semantic models for neurolinguistic decoding. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 114–123. Association for Computational Linguistics, 2012. 2.2.2, 4.2.2, 1

[86] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fMRI. *NeuroImage*, 56:400–410, 2011. 2.2.1, 6.2

[87] Willie Neiswanger, Chong Wang, and Eric Xing. Asymptotically exact, embarrassingly parallel MCMC. E-print, arxiv:1311.4780, 2013. URL `http://arxiv.org/abs/1311.4780`. 3.5.5

[88] Adrian Nestor, Jean M Vettel, and Michael J Tarr. Internal representations for face detection: An application of noise-based image classification to BOLD responses. *Human brain mapping*, 34(11):3101–3115, 2013. 4.4

[89] A Newell. You can't play 20 questions with nature and win. *Visual information processing. New York: Academic Press*, 1973. 1, 1

[90] S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J.L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 2011. 1, 2.3, 2.3.1

[91] S. Nishimoto, A.T. Vu, T. Naselaris, Y. Benjamini, B. Yu, and J.L. Gallant. Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, 2011. 7.1.1

[92] J. Nivre, J. Hall, J. Nilsson, A. Chanev, G. Eryigit, S. Kubler, S. Marinov, and E. Marsi. Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering*, 13(2):95, 2007. 4.2.1, 4.2.2

[93] Kenneth A Norman, Sean M Polyn, Greg J Detre, and James V Haxby. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in cognitive sciences*, 10(9): 424–430, 2006. 2.2.1, 6.1

[94] M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell. Zero-shot learning with semantic output codes. *Advances in neural information processing systems*, 22:1410–1418, 2009. 3.2

[95] C. Pallier, A.D. Devauchelle, and S. Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, 2011. 4.4

[96] D.G. Pelli. The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, 10(4):437–442, 1997. 4.1

[97] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, 45:S199–S209, 2009. doi: 10.1016/j.neuroimage.2008.11.007. 3.5.4

[98] Danny Pfeffermann. New important developments in small area estimation. *Statistical Science*, 28(1):40–68, 2013. 3.5.2

[99] Russell A. Poldrack. The role of fMRI in cognitive neuroscience: where do we stand? *Current Opinion in Neurobiology*, 18:223–227, 2008. doi: 10.1016/j.conb.2008.07.006. 3.5.4

[100] Beatrix Potter. *The Tale of Peter Rabbit*. ABDO, 2006. 4.1

[101] Friedemann Pulvermüller. Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, 6(7):576–582, 2005. 2.1.1

[102] J.N.K. Rao. *Small Area Estimation*. Wiley, New York, 2003. 3.5.2

[103] N. Rao, R. Cox, C.and Nowak, and T.T. Rogers. Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. In *Advances in neural information*

processing systems, 2013. 6, 6.1

[104] Susan M Ravizza, Marlene Behrmann, and Julie A Fiez. Right parietal contributions to verbal working memory: spatial or executive? *Neuropsychologia*, 43(14):2057–2067, 2005. 4.4

[105] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological bulletin*, 124(3):372, 1998. 4.2.2

[106] J.K. Rowling. *Harry Potter and the Sorcerer's Stone*. Harry Potter US. Pottermore Limited, 2012. ISBN 9781781100271. URL `http://books.google.com/books?id=wrOQLV6xB-wC`. 4, 4.1, 4.3, 5, 5.3, 5.3.1

[107] Magnus Sahlgren. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. 2006. 4.2.2

[108] Riitta Salmelin. Clinical neurophysiology of language: the MEG approach. *Clinical Neurophysiology*, 118(2):237–254, 2007. 1, 2.2, 2.1.1, 5.1

[109] R. Saxe et al. Uniquely human social cognition. *Current opinion in neurobiology*, 16(2): 235–239, 2006. 4.4

[110] Bernhard Scholkopf and Alex Smola. *Learning with kernels*. MIT press Cambridge, 2002. E.1.1, E.2, E.4.2

[111] Steven L. Scott, Alexander W. Blocker, and Fernando V. Bonassi. Bayes and big data: The consensus Monte Carlo algorithm. Presented at the "EFaBBayes 250" conference, 16 December 2013, Duke University, 2013. 3.5.5

[112] Gordon M. Shepherd. *Neurobiology*. Oxford University Press, 3 edition, 1994. 3.5, 3.5.3, 3.5.5

[113] Alex J Smola and Bernhard Schölkopf. *Learning with kernels*. Citeseer, 1998. 6.1, 6.2

[114] N.K. Speer, J.R. Reynolds, K.M. Swallow, and J.M. Zacks. Reading stories activates neural representations of visual and motor experiences. *Psychological Science*, 20(8): 989–999, 2009. 1, 2.3.2

[115] Peter Spirtes, Clark N Glymour, and Richard Scheines. *Causation, prediction, and search*, volume 81. MIT press, 2000. E.1

[116] Charles Stein. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, 1(399):197–206, 1956. E.1.3

[117] Gustavo Sudre, Dean Pomerleau, Mark Palatucci, Leila Wehbe, Alona Fyshe, Riitta Salmelin, and Tom Mitchell. Tracking neural coding of perceptual and semantic features of concrete nouns. *NeuroImage*, 62:451–463, 2012. 2.1.1, 3.1, 3.5, 3.5.4, 5.3.2

[118] Gábor J Székely, Maria L Rizzo, Nail K Bakirov, et al. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794, 2007. 2

[119] Samu Taulu and Juha Simola. Spatiotemporal signal space separation method for rejecting nearby interference in MEG measurements. *Physics in medicine and biology*, 51(7):1759,

2006. 5.3.1

[120] Samu Taulu, Matti Kajola, and Juha Simola. Suppression of interference and artifacts by the signal space separation method. *Brain topography*, 16(4):269–275, 2004. 5.3.1

[121] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996. 3.5.1

[122] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, M. Joliot, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage*, 15(1):273–289, 2002. 3.5, 4.1

[123] Ashish Vaswani, Yinggong Zhao, Victoria Fossum, and David Chiang. Decoding with large-scale neural language models improves translation. 2013. 5.2

[124] Ulrike Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4): 395–416, 2007. 3.5.3

[125] Martin J Wainwright and Michael I Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008. 3.5.5

[126] Leila Wehbe, Aaditya Ramdas, Rebecca C Steorts, and Cosma Rohilla Shalizi. Regularized brain reading with shrinkage and smoothing. *Annals of Applied Statistics, in press*. 2, 3.5

[127] Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one*, 9(11):e112575, 2014. 1, 3, 4, 5.3.2, 6.3, 6.5

[128] Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014. 4, 5, 5.2

[129] A. Wu, M. Park, O.O. Koyejo, and J.W. Pillow. Sparse Bayesian structure learning with "dependent relevance determination" priors. In *Advances in Neural Information Processing Systems*, 2014. 6

[130] Tal Yarkoni, Russell A. Poldrack, Thomas E. Nichols, David C. Van Essen, and Tor D. Wager. Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, 8:665–670, 2011. doi: 10.1038/nmeth.1635. 3.5.4

[131] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320, 2005. 3.5.1