Machine Learning Methods for Computational Psychology

A Dissertation Presented

by

Sarah M Brown

to

The Department of Electrical and Computer Engineering

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in

Electrical Engineering

Northeastern University

Boston, Massachusetts

December 2016

Contents

List of Figures

List of Tables

List of Acronyms

BOLD Blood-oxygen level dependent.

DBN Dynamic Bayesian Network.

ELBO Evidence Lower Bound.

fMRI functional Magnetic Resonance Imaging.

FOV Feild of View.

GP Gaussian Process.

HRF Haemodynamic Response Function.

NIfTI Neuroimaging Informatics Technology Initiative.

NLL Negative Log Likelihood.

PGM Probabilistic Graphical Model.

UFBMM Unified Framework for Brain-Mind Modeling.

Abstract of the Dissertation

Machine Learning Methods for Computational Psychology

by

Sarah M Brown

Doctor of Philosophy in Electrical and Computer Engineering

Northeastern University, December 2016

Professor Jennifer Dy, Adviser

Advances in sensing and imaging have provided psychology researchers new tools to understand how the brain creates the mind and simultaneously revealed the need for a new paradigm of mind-brain correspondence – a set of basic theoretical tenets and an overhauled methodology. One emerging candidate paradigm is the Unified Framework for Mind-Brain Modeling (UFMBM). We develop complementary machine learning methods necessary to overcome initial barriers to adoption of the new paradigm. We assess candidate solutions to these problems using a novel dataset aiming to verify theoretical tenets of the new paradigm in a study of basic affect using functional Magnetic Resonance Imaging (fMRI).

Re-analysis of existing datasets in light of the proposed theoretical tenets is a convenient method for demonstrating the power of the new paradigm. These datasets may be statistically small relative to the new class of models, so differentiating between a weak signal and an overfit model is an important capability for promotion of the new paradigm. We relate theoretical results for stability as a guarantee for good generalization to its empirical application as a performance measure in unsupervised learning in order to apply metrics derived from this template to subsequent analyses.

We begin to empirically evaluate the theoretical tenets of the new paradigm in an extended fMRI study, with 900 trials. As a baseline, we apply standard analysis methods on increasing lengths of data to demonstrate widespread engagement in task. We propose Bayesian nonparametric models for structure discovery under uncertainty to model the mind-brain relationships of interest. The proposed model learns groupings of voxels based on shared response shape characteristics and groupings of trials based on shared spatial activation patterns. The model does not require a priori specification of the number of groups of either type due to Dirichlet process priors on the partitions and shares information across non-uniformly sampled trials through a Gaussian Process prior on function shape.

Chapter 1

Introduction

Scholars have tried to develop an understanding of the human mind by studying the brain for over two thousand years. Despite this long tradition, an understanding of how the brain creates the mind is still one of the greatest challenges in science, and is broadly funded by major research efforts across the globe [Grillner et al., 2016]. Discovering how the brain works with sufficient granularity to understand the mind is not only a great intellectual pursuit, it holds immense implications for society because grounding our understanding of human behavior in the physical brain enables interventions. For example, an understanding of learning could improve education, or an understanding of those problems underlying mental pathologies could improve care. Machine learning and cognitive science have shared a long, collaborative history. Numerous machine learning methods are based on hypotheses about the brain and computer-based analogies have informed psychological explanations [Marsella et al., 2010]. In this work, we focus on experimental data analysis–using data that was collected to learn about human brain-mind mappings–rather than mimicking behaviors.

As measurement tools become increasingly available, more sciences can proceed in a data driven manner. In order to explain and synthesize this data, psychological research requires new theory and corresponding empirical analysis techniques. In psychology research, neuroimaging and other sensing technologies present a new opportunity for psychologists to study how the brain creates the mind, enabling a quantitative exploration of those classes of questions that have been historically limited to philosophy. Unfortunately, interpreting this data pushes the limits of the existing computational techniques popular in behavioral sciences. This is best suited to machine learning techniques that make sense of data by finding patterns, revealing latent structure and generating predictions.

Recently, neuroimaging studies have produced a series of results that challenge the core theoretical tenets of the dominant paradigm in psychology research. In order to accommodate this new experimental evidence, psychologists must adapt the theory, experimental design and analysis techniques they most commonly use; psychology is in need of a paradigm shift [Spivey, 2008, Barrett, 2017, Barrett and Satpute, 2013, Uttal, 2001]. We depend on a new paradigm that is presented in three parts: theory, model and algorithm. The theory comprises six key tenets, each of which identifies a departure from essentialism. The theory also addresses sources of variance unacknowledged by faculty, common sense psychology, current experimental design and current methods of data analysis. The theory informs an abstract, ideal, mathematical model–one that is far too complex to be implemented directly for data analysis. Theoretical concepts are translated into mathematical properties of the model. Finally, the framework adopts an algorithmic approach to the model.

The Unified Framework for Brain-Mind Modeling (UFBMM) calls for probabilistic methods to directly model uncertainty in experimental data. We take as a point of inspiration the grand challenge to design machine learning methods that can facilitate the transition from one paradigm of psychology research to another. All disciplines of science move through phases of puzzle solving and crisis. Puzzle solving is a period where a shared paradigm sets up common theories accepted by (nearly) all in the field with a corresponding set of questions, and scientists seek answers to these questions. Crisis is when the results no longer match the theory, and a new paradigm, core theory and accepted questions are necessary [Kuhn, 2012]. Old theoretical ideas cannot simply be abandoned, they must be replaced. Replacement requires new theory and corresponding empirical evidence for the new ideas. To accumulate new empirical evidence, old analysis methods reflecting the old theory must also be replaced.

In this dissertation, we consider the implications of this paradigm shift in the context of studying affect with functional Magnetic Resonance Imaging (fMRI). Our goal is to both to demonstrate the new paradigm's utility, and move toward better designed experiments. We define two specific technical barriers to analyzing experimental data and test our approaches in synthetic and experimental data, as illustrated in Figure 1.1. First, we ask questions that are categorically different from prior results. Because we specifically aim to test computational techniques derived from hypotheses that are ahead of current experimental design strategies, there are no performance metrics inherent to the problem domain. To address this, we propose stability, which has been used in machine learning in other contexts, as a class of performance metrics that are robust to data size and adaptable to this application. Second we integrate Bayesian nonparametric modeling with standard fMRI analysis techniques to allow for flexible structure discovery, while gradually

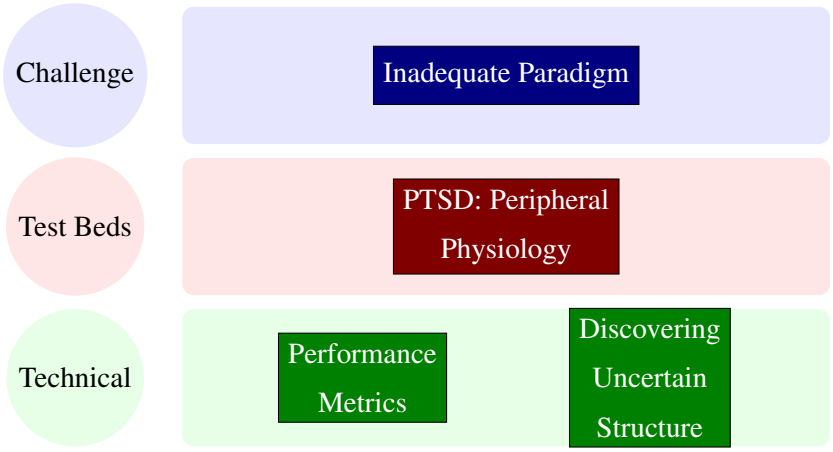


Figure 1.1: Project over-view in terms of grand challenges, test beds and technical barriers.

relaxing assumptions made in analysis. We demonstrate and assess results in synthetic data, and further evaluate performance in an fMRI dataset that studies core affect.

The main conclusion of this thesis is that machine learning derived approaches can provide the methodology necessary to advance psychological research by managing data and creating empirically validated theories of how the brain works. Specifically, we present three distinct insights driven by in machine learning. First, algorithmic stability is a technically principled and interpretable performance measure for highly uncertain data analyses in domains where predictive accuracy misses key information. Second, physiological correlates of behaviorally defined diagnostic scores may be learned by defining a novel cost function. Third, in combination with relaxed assumptions in the experimental design, Bayesian nonparametric techniques for structure discovery enable neuroimaging analyses to be more data driven, while still sufficiently robust to over-fitting to allow valid scientific conclusions.

In the remainder of this chapter, we further develop the problem statement with respect to the application area, providing context and motivation. We foreshadow technical challenges which will be detailed in the next chapter, section 1.4, by providing shared technical background material. We use the structure of Figure 1.1 as an organizing principle. In section 1.1 we review the broad challenge of understanding sensor data in psychology in an inclusive, consistent way. We describe the specific test bed applications in section 1.3, but we leave the details of the data, as prepared and provided to the algorithms to the technical chapters. Finally, we summarize the specific technical challenges and the novelty of contribution with respect to each in section 1.2. The technical chapters (chapter 2, and chapter 3) detail each challenge and review fully the technical literature with respect

to each of these problems.

1.1 Grand Challenge: the human mind and brain

Psychology has been in a crisis for a number of years. As many other fields of science have experienced in the past, psychology is on the verge of taking a leap away from an essentialist approach and the current dominant paradigm–faculty psychology–toward a more inclusive view of variance. With collaborators we propose a new framework consisting of theory, a model, and an algorithmic methodBarrett et al. [2017a]. We identify six key phenomena for which the current paradigm does not provide good explanations and provide an alternateBarrett et al. [2017b]. Instead of addressing problems with current analyses or individual hypotheses and modifying standard analysis procedures to accommodate, we propose a ground up replacement of the process. We propose a new framework based on idealized model of the brain and how it creates the mind. Subsequently, this lays the ground for a joint research agenda in machine learning and psychology that develops the tools to analyze data collected with experimental designs motivated by this framework. In this thesis, we take a small step on this agenda. For framing, we provide the full description of the framework here, noting that its complete implementation is beyond the scope of the thesis.

As is common with scientific redirects, technology is a necessary impetus for progress, but in our current data-intensive state of science, improved or even new measurements through instrumentation are no longer sufficient. Today, these advances also require novel computational approaches in order to interpret the measurements in the context of the theory. Neuroimaging and better sensors for peripheral physiological have been used to try to understand how our brain creates the psychological experiences we have and how our bodies react to them. To date, the results have been largely inconclusive, or even inconsistent. The new measurements have revealed flaws in theory rather than provided further confirmation. Current methods for interpreting the data and relating it to hypotheses are a barrier to understanding the flood of new data being generated.

The UFBMM is a paradigm for psychology research that explicitly states distinct psychological theory, analytical modeling, and algorithmic techniques. This explicit distinction, combined with the specific theoretical tenets it supposes addresses many of the apparent contradictions between recent results in neuroimaging and the current common-sense faculty psychology paradigm[Barrett et al., 2017b]. The theory posits six key tenets: 1. the *whole brain* engages to create mental experiences, which vary as spatio-temporal patternsDeco et al. [2015] 2. every network, region, and neuron activates in multiple psychological purposes; the brain is *domain general*,

3. the brain has multiple ways of creating the same perception or psychological process; like other biological systems the brain *degenerate* 4. past brain activity influences the present, 5. the brain is an always on *dynamical system* with memory 6. inputs to the brain are complex and include all surroundings including the body and that experimenter bias needs to be modeled and finally. Based on the need to accommodate these criteria, the UFBMM model defines the latent variables in terms of their meaning and the necessary mathematical properties in order to capture the necessary phenomena. The framework suggests a probabilistic approach to analyzing data with the model in order to incorporate the various sources of uncertainty in leveraging data to verify and extend the theory. Given our objective to develop a model for analyzing neuroimaging data, the following presentation of the theory will be anchored in the modeling requirements. We will present the model as a set of definitions, and anchor the definitions in probabilistic modeling.

1.1.1 Unified Framework of Brain-Mind Modeling- Theory

Central to a theory of brain-mind mapping is a theory for how the brain creates a mental experience. The brain is comprised of billions of neurons and glial cells interacting through electrical activity flowing along axons and neurotransmitters signaling between them. Despite their extreme number, the whole brain works as a unit, to create mental experiences. This stands in contrast to the common sense view of psychology that looks for spatially localized regions to increase activity when its concept is perceived. Since the whole brain works together to create each mental experience, that also means that every part of the brain is involved in many psychological processes, referred to as domain generality. Third, psychological processes can be created through multiple neural mechanisms. These ideas are complementary, but each is supported by distinct evidence.

The evidence that various individual regions engage in multiple psychological ideas has been occasionally attributed to the idea that the previous labeling was wrong. However over time enough evidence has mounted that it is no longer likely that is the case. Additionally, work looking to isolate various psychological ideas to a single portion of the brain has been upended by running longer studies, by adding more stimuli, or examining more activity. The idea that the brain is degenerate means that there are multiple pathways a brain can carry out a given function.

The brain is a dynamic system and mental experiences are dependent on the brain state. Many psychology experimenters treat the brain as a stimulus and response system. This treatment is technically equivalent to an assumption of a memoryless system; there is no modeling of how past neural activity impacts present activity. There is, however, evidence in limited cases that this is a

necessary approach. Similarly, these events should not be modeled as additive to background activity because brain activity is continuous rather than stimulus driven, and psychological information is held in these time-summarizing states. Also even in a controlled experimental setting, the inputs to the brain are not within control of the experimenter. Recent studies have also shown that various brain functions are modulated by other parts of the body, for example circulating glucose levels, in a variety of ways.

1.1.2 UFBMM- model

We propose modeling the the brain as a dynamic system, with a brain state $\mathbf{b}(t)$ and brain state dynamics function $B(\mathbf{b}(t), u(t), t)$ where $u(t)$ represents the inputs to the brain. The brain state includes all physical quantities in the brain (electrical, chemical, etc.) and, in order to be a proper state vector, their temporal derivative information. The inputs to the brain are all quantities exogenous to the brain, that can influence the brain state's trajectory, including both designed stimuli and the influence of the body. The brain state dynamics function, B must be time varying. We say that the brain state lives in a space $\mathcal{B}$.

In order to study psychology, we need to introduce a representation of the mind, in particular we choose to say the mind exists in a space $\mathcal{M}$ that is related to the brain through a lossy projection g. At a given moment in time an individual's mental field $\mathbf{m}(t) = g(\mathbf{b}(t))$ defines their entire mental experience, both conscious and unconscious. To be able to best address the criteria described above we introduce two types of psychological concepts for labeling the mental field. First, mental features are psychological primitives that define the directions of the mental space, $\mathcal{M}$, denoted by $m_i(t)$. What precisely these are is an active area of debate in psychology, with hope that this framework can provide a framework for empirical discussions. Second, mental categories are broader psychological concepts, defined by a region in the mental space, $\mathcal{M}$. Mental categories are not continuous and can overlap, thus the operation $c(\mathbf{m}(t))$ returns a list of discrete variables of varying length depending on the value of $\mathbf{m}(t)$. An example mental category is "anger". To say that a subject is angry means that the brain state is in the pre-image of the "anger" region of the mental space. Since the mental category can be noncontinuous, there could be two brain states that are both physically distant and have very different mental descriptions (values of each mental feature) that are both anger. Because mental categories can overlap, a brain state can be both anger and sadness.

To complete our model, we also describe physiological features, a lower dimensional projection that serves as a latent intermediary between the physiological measurement and the brain.

Physiological features are defined through the function f, where $\mathbf{x}(t) = f(\mathbf{b}(t))$. Those physiological features that are of interest for analysis of a specific experiment depends on the choice of measurement. Finally we introduce measurement operators. In line with the necessity to model all inputs, it is imperative to explicitly model the measurement system. We can measure both physiological $\mathbf{y}_x(t) = \mathbf{Y}_x(\mathbf{x}(t))$ and mental $\mathbf{y}_m(t) = \mathbf{Y}_m(\mathbf{m}(t))$ features.

We also know that g, the brain to mental field projection, is partially invertible (i.e., some portion of the brain state could be recovered from a complete observation of the mental field). Correspondingly, f is also invertible. The pre-image of $\mathcal{M}$ and $\mathcal{X}$ overlap in $\mathcal{B}$, but not completely; this is an important aspect to maintain in our modeling efforts. For any choice of f there is information in the mental field that is not in the physiology and there is also information in the physiology that is not in the mental field. Therefore, for any given set of behavioral and physiological measurements, there is meaningful variance in each that is not captured by the other. Of course, experimental design aims to choose measurements such that they overlap largely, but designing models that accommodate this type of uncertainty explicitly is important as well. Given these definitions, we can frame the necessary mathematical properties using the theoretical tenets they help us address. Psychological phenomena are not localized; our model must consider the whole brain as a unit with respect to the mind, not as pieces. For any given mental feature, $m_i = g_i(b_i(t))$ the subset of b in b_i is always a spatially global representation, though it may depend on different areas to varying degrees over time. We cannot decompose the mental field such that $\forall i, m_i = g_i(b_z(t))$ where $z \in R^3$ refers to a spatial location, region, or network and m_i is a single mental feature. Therefore mental categories, as regions in the space $\mathcal{M}$, cannot be spatially localized in the brain either. Brain structure by itself does not map one-to-one to high level descriptions of psychological function, and perhaps not to lower level psychological features. Psychological function depends on trajectories over time and not just on static structures; models must account for this. Mental categories, like fear and anger, are overlapping regions in the mental feature space, the operation $c(t) = C(m(t), t)$ returns a varying length list of discrete variables, where C is not uniquely invertible and $m = g(\mathbf{b}(t)$. For any spatial region in $\mathcal{B}$ the image in M is high dimensional. That is every $b_i(t)$ is in the domain of multiple g_j where $m_j = g_j(b_j(t))$. Brain activity is continuous in time; a model must be dynamic and include descriptions of temporal dependencies, b is a state vector and B is a time varying description of the dynamics. The mind depends on spatio-temporal patterns of brain activity; a model must account for the influence of brain dynamics on the mind, $m = g(b)$ where b is state vector that contains temporal information, and g depends on more than just the components of b that represent instantaneous activity. Though g is an instantaneous mapping, it only depends on the temporal information

already contained in the state vector, not on its own history. The mind reflects the brain's response to its total environment, including both exogenous and endogenous quantities. In particular, influences beyond the stimulus are psychologically potent; a model must explicitly consider all influences on the brain. Multiple, distinct brain states (spatio-temportal patterns of electrical and chemical activity) may have a single mental label. A model must allow many to one mappings from the brain state to the mental field.

1.2 Technical Barriers

The distinction of theory, model, and computation in the UFBMM presents great opportunity for machine learning research. Machine learning takes steps toward modeling uncertainty in data and extracting information from massive data sets and small data sets alike where uncertainty abounds. The challenge exists in working with new types of uncertainty and introducing the practical constraints inherent in working with a *real* problem. Experimental research creates constraints and outcomes that are different than working with data that will go into a commercial product. Reproducibility of results is one major concern in pilot scaled data sets, but is generally essential to progress in human subject fields.

Given this common technical framework, we identify three key challenges. First, practically, in order to advance science in this mode of operation, we need to interpret and assess performance of methods in non-ideal environments. These mimic environments where stability has served as a successful heuristic, but we would like to operate with some more detailed understanding of what these metrics mean. Next, our theoretical assumptions directly question the validity of what might serve as labels in a supervised learning setting. In these contexts, we need formulations for learning from ambiguous labels and that incorporate specific domain knowledge. Finally, in order to move toward implementations of the model for future data analysis, exploratory analysis of structure discovery in the data is an important step toward data driven formulation of hypotheses.

1.2.1 Unsupervised Model Fit

First, we consider stability as a candidate for a general framework for measuring performance in machine learning applications. We consider the literature that is there, determine an appropriate common definition and means to using stability as an assessment framework from which a researcher can easily design unit- and context-appropriate performance measures without need-

ing to determine a new theoretical interpretation or guarantees for the metric. We have crafted the general framework and the theory that interprets those so that new methods can be derived and only held to a few simple checks to determine if the theoretical properties hold. The literature review here includes how we define new measures.

1.2.2 Nonparametrics for uncertainty in fMRI

Standard fMRI analyses make a series of simplifying assumptions in order to glean out a signal. However, in light of a new theory for the key ideas, many of these assumptions should be relaxed. These assumptions are no longer safe numerical simplifications, averaging out errors of various sorts, but now detrimental oversimplifications, erasing meaningful variation. The new theory, does not however, automatically provide new ways of simplifying the problem to reach a computationally tractable point. Instead, we must innovate in analysis methods. We need to more explicitly model uncertainty instead of simply averaging it away.

For example, instead of assuming there is a fixed response each time a region (a voxel includes many neurons, but it is a small region) fires, we might learn from the data what a response looks like in a given region. Instead of assuming that the dominant similarities in patterns of response are the property of the stimulus the experimenter is most interested in, we might learn which trials have the most similar neurological response and then post-hoc determine the shared stimulus properties. Instead of assuming that at each time there is a single response and noise, or run off of a previous, we might allow multiple processes to co-occur.

The standard model used in neuroimaging analysis is a generalized linear model (GLM). This model assumes that the observed time course, y is the product of a design matrix X and a regressor vector β. The design matrix is constructed as one column for each experimental condition, a set of columns for nuisance variables and a constant for the mean. The design matrix encapsulates our model for what a 'response' to a stimulus is, as a set f basis functions in the condition columns. The nuisance vectors allow us to 'filter' out the portion of the signal that is better explained by those than by the expected model. By multiplying with one weight for each condition for each voxel, we learn how much each condition contributes to each voxel. The most common model for a 'response' is that neurons that encode something present in the stimulus will fire at an increased rate in response to the stimulus. This increased response may be either a delta at the onset or a step for the duration. Then, due to the increased firing, blood rushes to the region, and the fMRI measures Blood-oxygen level dependent (BOLD) signal, sensitive to this increase in hemodynamic flow. This

is modeled by convolving the 'neural' response with a Haemodynamic Response Function (HRF). A typical HRF is a double gamma.

Each condition column is typically constructed with an impulse at the onset of or a step function for the duration of the stimulus presentation, convolved with a hemodynamic response function. This is to model the influx of blood in response to increased firing. This model represents the idea that there is a spike (or box car) of increased firing by neurons in the voxels that correspond to the condition and that that increased firing requires an increase in blood.

This model makes a lot of assumptions. The typical model, learns the beta regression weights and then uses a significance test to produce spatial maps. The maps here are hard to interpret, the betas themselves should be maps.

Accommodation of time in understanding psychological processes is a core value of the new paradigm and thus time series modeling through state spaces is essential to the later parts of the thesis. In the state space modeling section we examine state space representations that include both the systems view and a probabilistic view. With both representations together, we can summarize results more fluidly and most importantly we are able to understand how the two representations interact. By comparing methodologies for approaching the same problem (state space modeling of time series data) we can see the advantages of each and combine and borrow from both to design more effective solutions. We also consider other methods of time series modeling: more traditional methods as well as flexible Gaussian process based methods. With all of these together we draw a conclusion of a means to converting between representations methods and how to mix and match the best solution methods. Models such as additive Gaussian Processs (GPs), etc allow rich relationships among complex functions and we can then integrate ideas like state space modeling and derive these concepts and interpret them within the scope of the application.

In order to analyze this data in a manner more reflective of current neuroscience knowledge, we need technical tools that are more reflective of how much is uncertain. Point estimates do not allow for that, they only reflect the peak. P values only represent a very weak claim that that there is some signal there, we wish to make more inclusive models. We propose Bayesian methods because they allow us to model the uncertainty directly, and maintain uncertainty from one step of analysis to another. Bayesian nonparametric models further allows us to capture a particularly important property of analyzing data from such a complex system as the human mind: the uncertainty in size. We know that within an experiment there is some finiteness, but we do not know where the cutoff is. If we collect more data, nothing says that it has to align exactly to something we have already seen. In fact, as people are always learning, it is quite certain that after some time, their

brain would create new patterns of activation. Why should we fix a certain model size based on our (as the experimenter) perception? Bayesian nonparametric models allow us to model discrete events as having an infinite number. While, a discrete model may not be the ideal model of the brain, some things will repeat and the amount of information lost in measurement is non trivial as well, so a discretization is not only convenient, but reasonable, but that we know a priori, or that every subject, even inn the same protocol goes through the same number is not necessary.

1.3 Functional Magnetic Resonance Imaging of Affect

Of course, the application of the new paradigm is a large body of work. Properly framing it defines many new questions reframes the interpretation of existing studies.

In this data set, five subjects, viewed a total of approximately 900 IAPSLang et al. [1999] images each, over five sessions while in an fMRI scanner. Each subject saw the images in a different order and no subject was shown any image twice. Within each session, the time was split into three to eight runs of approximately eight minutes, collecting 164 whole brain volumes with a TR (sampling rate) of 3s. Each image was shown for 4s followed by 10-14s (uniformly randomly distributed) of fixation, a + sign. The fixation time was varied to prevent a response that was primarily due to the subject anticipating the image. To keep the subjects attentive, some of the fixations were red instead of black and the subjects were told to press a button on the red fixations. The images were selected to be of positive, negative and neutral valence and high and low arousal as per the normed ratingsLang et al. [1999] and randomized to have a useful distribution of the transitions among those and of consecutive presentation so as to compare how many back a modulation factor of the previous image influencing the present image may last. Outside of the scanner the subjects rated the images for valence, positive or negativity, and arousal, or intensity.

For the all results, we follow a standard fMRI preprocessing stream, described in detail below. All preprocessing was managed with NiPypeGorgolewski et al. [2011] and completed using FSLJenkinson et al. [2012].

1.3.1 Rotate structural to standard space

The functional scans were collected in close to standard orientation, but the anatomical scans were not. For ease of understanding and standardization, the anatomical scans were rotated to

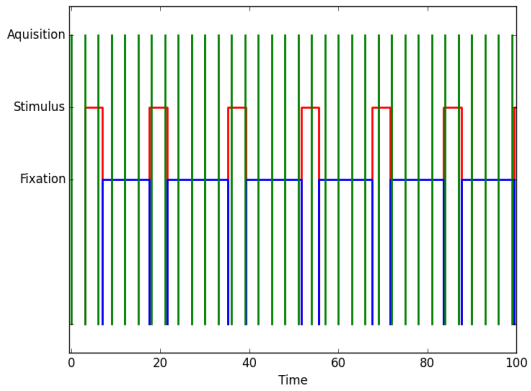


Figure 1.2: Illustration of the timing of a segment of the experiment. Green spikes represent the aquisition times for the fMRI volumes (TR=3s). The red bars show the stimuls presentations each 4s, and the blue bars show the fixation time, $U(10, 14)$.

standard orientation with the FSL tool `fslreorient2std`Jenkinson et al. [2012] at the command line for each subject.

1.3.2 Numerical Preparation

After the data was converted from DICOM to Neuroimaging Informatics Technology Initiative (NIfTI), the sub-sections of each run were merged in time to produce a single NIfTI file per run with `fslmerge`Jenkinson et al. [2012]. At the start of the preprocessing script executed with NiPype, the files were converted to floats from integers, to avoid integer division.

1.3.3 Slice Time Correction

To account for the time delays within the collection of a single plane of the image, the slice timing correction was applied using FSL `SliceTimer` with interleaved setting as True. This processes each voxel independently and uses Sinc interpolation with a Hamming windowing kernel to 'align' all slices within each volume.

1.3.4 Realignment

Each volume within a run was linearly transformed to align with the center volume of that run using FSL MCFLIRT Jenkinson et al. [2002],. This adjustment corrects for small amounts

of motion. The amounts of correction were saved and reviewed to check for large amounts of displacement and correction matrices for use as nuisance regressors in analysis.

1.3.5 Skull Stripping and Anatomical Segmentation

Using BETSmith [2002] we extracted the brain from both the anatomical and mean of the aligned functional in order for the registration to work properly, with default fractional intensity threshold of 0.5. For the anatomical skull stripping we use the -B setting, which removes bias and residual neck. Without this some runs do not register properly.

Anatomical Segmentation of subcortical structure with FSL FIRSTPatenaude et al. [2011].

1.3.6 Registration

We register the motion corrected functional scans for each run to the subject's structural, but do not register to a standard brain. The skull-stripped motion corrected functional mean is registered to the skull stripped standard orientation anatomical scan and the linear transformation matrix is saved. Registration is performed with FSL FLIRT Jenkinson et al. [2002], Jenkinson and Smith [2001], Greve and Fischl [2009] and Correlation Ratio cost function(default), seven degrees of freedom(6 for rigid body, 1 additional because the voxels in the structural and functional are different sizes), and spine interpolation (in the apply step). The transformation is first applied to mean for use later in the masking step, with a manually created reference image file used for voxel size and FOV. Then the transformation is applied to each aligned functional run, (without skull stripping) using the manually created reference image for voxel size and FOV. The reference images is required because otherwise interpolation (keeping the voxel size) from the functional to the anatomical is also applied.

A single reference image for Feild of View (FOV) and voxel size was made for the whole study (shared across subjects and runs) at the command line by getting the header information from both a functional scan and an anatomical image with `fslhd`. The field of view size in a copy of functional header were manually changed to be proportionally as large as the anatomical scan, given larger voxels. Then an empty reference image was created with `fslcreatehd`. This process was derived based on the FLIRT FAQ"Analysis Group at FMRIB" and FSUTILs guideat FMRIB".

1.3.7 Masking

First, we construct a mask for each run, then take intersection across all runs for each subject as the overall subject mask. This ensures that the same voxels are included in all runs. This procedure is that same procedure as the Gonzalez-Castillo et al. [2012] mass averaging study. To create liberal functional masks, we apply FSL `BET`Smith [2002] to strip the skull from each warped functional mean with fractional intensity threshold set to default of 0.5. Next we apply `FSLstat` to determine the intensities at the 2nd and 98th percentiles of each functional run and dialate the mask. This mask is applied to the data before subsequent analyses. For use in analyses, we also create a tighter brain mask, by registering the anatomical mask created in the brain extraction step above to the output of the registration of the functional to anatomical. In this case `FLIRT` is used to resample the anatomical mask and `fslmaths` is used with threshold 0.5 to binarize the image.

1.3.8 Spatio-Temporal Smoothing

Spatial smoothing was applied with SUSANSmith and Brady [1997], a Gaussian filter with FWHM of 5mm. Hihg pass filtering removes low frequency content that is attributed to MR physics not brain activity. We use a cutoff of 100s set as FWHM in of (100/(2*3) =) 16 volumes.

1.3.9 Normalization

Normalization is required to compare across runs because the fMRI data is unit-less and varies run to run due to a lot of uninteresting factors. We normalized the data by dividing each run by its spatio/temporal median.

1.4 Technical Approach and Background

The central thesis is that machine learning provides a useful set of tools for psychology to overcome this period of crisis; to move away from the old paradigm and begin to produce empirical results derived from the UFBMM. Novel data analysis can fuel data driven discovery, in this chapter we frame how machine learning can work on these problems and provide the core technical background and constants. Machine learning is often viewed in two ways, model based or toolbox based. We advocate for the former in the context of data driven discovery; pulling algorithms out of a 'toolbox' and evaluating results could easily lead to misuse of the data in context and invalid reasoning.

We need new methods and we want to incorporate flexibility and uncertainty inherent in scientific discovery. This is the gateway to this application of machine learning pushing the boundaries of machine learning ability. Given this, we are advancing science and human understanding, we need to restrict our machine learning solutions to models that are interpretable to humans. Our algorithms can be complex, but the returned model is to be an explanation of the data, it needs to be translatable into a story, not a black box for associating inputs to outputs. An empirical prediction given an observation is not sufficient, in science, we need a model that allows us to generate new hypotheses. In this sense, we can liken machine learning for science to a longer time scale that a typical machine learning application. Science is an ongoing process, the scientist is in the loop of the computational learning and we want a prediction for a whole experiment over the scope of potential new experiments, not just within a single data set. We need to be able to make predictions over hypothetical datasets that have not yet been collected. This restricts us to the case of generative models, as opposed to discriminative models.

The core idea is that machine learning methods can facilitate refining scientific theory by integrating empirical evidence in more robust ways than simple statistical methods traditionally employed in science. This implies that machine learning provides an assist at a philosophy of science point of entry. This idea is not new, there is existing work that proposes machine learning as a form of philosophy of science and many 'empirical philosophers' are closely aligned with machine learning. There are related processed in the two fields, but they remain complementary, not to be merged[Williamson, 2004, 2010]. There are separate levels of analysis, though. What we will call the 'outermost' level is that of choosing one model over another- commonly hypothesis testing. Developments in Bayesian learning through machine learning research support this endeavor. Deeper than that, however is the modeling. A hypothesis test assumes that there must be something- but what do you form as a hypotheses in the data? Simple analyses such as an increase in the mean, no longer provide a granularity of detail that advances many sciences. It is now necessary to express richer hypotheses and to infer properties and validity of these from data. This is where machine learning excels most. However, by retaining that our decision is at the outer level we can incorporate the structure of the two steps in more efficient and interpretable ways.

For the purpose of explanatory power, we will define interpretability as a model that can be expressed concisely at least through visualization, if not an analytical form. For a predictive model, like a diagnostic scoring function, interpretability means that a model must be simple enough to a clinician to understand what the function is doing. In some work, interpretability is applied to mean that latent factors and a clear meaning that can be communicated to a person. For example

an interpretable clustering of food pictures might be sweet versus savory or categories of foods. That the latent concept has a meaning that is commonly used. In our case, since we are aiming for discovery, we do not require interpretability in this sense, only in that we have relationships between the latent and observed that are useful. For this reason, we will introduce a requirement for model-based methods over black box methods and probabilistic relationships instead of strictly additive noise- thus preferring Bayesian inference.

We adopt pragmatic Bayesian interpretation of probability. That is, in general, we have Bayesian interpretations, that a probability is a belief, and we will make use of priors, in general. We interpret probability as a quantification of uncertainty, not as the limit of infinite events However, some parts of the model may not be as important as others and so we may adopt frequentist methods as approximations in some cases, such as when a valuable prior is not available or easily selected. Reasonable approximations that mix Bayesian thinking with frequentist techniques provide convenient computational advantages and escape the pitfalls of frequentist methods that make for dangerously over-assertive solutions. For example, an easy maximum likelihood type solution may be sufficient - in other cases that simplification may miss important error. Instead of a hard methodological decision, we take a case by case approach using context to choose an algorithm- truly being in a way even more Bayesian- in a meta-sense. We incorporate non-empirical evidence inn our decision making about the outer-problem of choosing inference. For more complex models a fully Bayesian inference method may be necessary for finding a solution that is adequate in terms of precision and flexibility. Expressing a valuable probability distribution over possible answers.

get cites

In order to define concisely the specific technical barriers to answering these broad questions and testing in the data as described that will be overcome by this thesis, we introduce the preliminary background that provides the framing. We consider a new model for computational psychology- using this framework approach shown in Figure 1.3, we also consider what defining properties exist for how to abstract a problem into a learning problem. In the UFBMM, we propose an algorithmic approach that explicitly models uncertainty, leveraging probabilistic graphical models as a language for expressing the model and acting specifically. In this section, we define a learning problem and delineate notation that will be used throughout.

First we address the core technical background and definitions we need in order to frame the main contributions of the dissertation. Further we address some points which seem to be at odds, but for an application actually balance well as complementary techniques. We take a pragmatic Bayesian approach to modeling. In the time series analysis, we focus on the application of Bayesian non parametric methods. However, in addressing the performance of the models, we begin

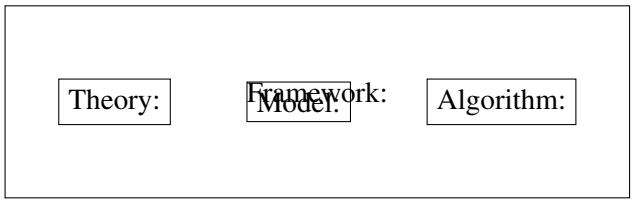


Figure 1.3: Caption

from a statistical learning theory perspective, which is generally routed in frequentist methods and assumptions. Though we aim to use Bayesian methods to finally conduct analyses; we are working in a domain where there is more uncertainty that there are known, some preliminary steps leveraging other methods are realistically useful for determining initial starting points and making decisions that enable the development of the more idealistic models. In a real data application, we must make compromises at some point; we use this set of analyses in order to help understand the implications of the compromises we may make.

In the UFBMM, we propose that probabilistic graphical models are well suited to implement analysis techniques derived from the model we have described in order to maintain the necessary uncertainty and flexibility for a reasonable model. Machine learning methods are well suited to this objective, the core objective is to find and extrapolate patterns in data. Here we provide the basic machine learning terminology and definitions necessary to understand the remainder of the thesis. Detailed technical literature reviews are left to the individual technical chapters, but here we provide common notation and conventions. Each chapter will have its own variable definitions, here we provide only conventions and conceptual definitions.

In general, machine learning aims to use examples of data to learn an underlying model.

1.5 Mathematical notation

A probabilistic graphical model is a visual representation of a joint probability distribution that consists of nodes and edges. Visually nodes are denoted by shaped and edges by connecting lines or arrows. Conceptually, nodes represent random variables and edges represent conditional dependencies. An arrow represents the direction of the conditional distribution, *i.e* P(a,b) = P(a—b)p(b), the joint of a and b is a given b times prior of b would be drawn with an arrow from b to a. Throughout we will use shading to denote an observed variable and no shading to denote hidden variables. For fixed parameters or hyper parameters, we will use a solid dot. Plates, or large

rectangles around a node or group of nodes will denote a repeated structure, where a number in the bottom right indicates the number of repetitions.

Bayesian approaches presume that parameters of the distributions are also random variables. Analysis is then inference of the posterior distribution of the unknowns, latent variables and parameters, given the observed data and hyperparameters that are assumed to be fixed.

We use both sets of methodologies to address different aspects of the problem at hand. While a model based approach leveraging, Bayesian nonparametric approaches is good for designing an eventual model, many domain scientists are accustomed to other approaches and to promote a real impactful machine learning solution compromise in methods that are interpretable is necessary. Further, in initial steps of a project, defining a good probabilistic model for the problem can be challenging. A quicker result, if only preliminary can be derived in a traditional statistical sense. Finally, many empirical risk or frequentist approaches can be derived as approximations from Bayesian models. Borrowing ideas across the two is a practical consideration and pathologies of frequentist approaches can be avoided by adhering more toward a model based approach, but allowing harder approximations when advantageous.

1.5.1 PGMs

Probabilistic Graphical Models (PGMs) are a compact visual representation of joint distribution among a group of random variables. They are designed to be able to clearly express out the factorized form of the distribution as a product by emphasizing the conditional independencies. In a PGM shapes, called nodes, are used to represent random variables and lines, called edges, represent probabilistic dependencies. Shading is used to denote observed variables or measured quantities. Larger boxes, called plates are used to denote repeated portions of the graph, with the number of repetitions marked in the bottom right of the plate. A deterministic constant is denoted by a dot node. There are a variety of classes of PGM, including directed graphical models or Bayesian Networks, undirected graphical models, and factor graphs. To model dynamic systems, we use a special type of Bayesian network, called a Dynamic Bayesian Network (DBN), this type of model encodes only conditional independencies and has an implicit sense of time[Murphy, 2002].

Scientific progress is not a smooth, continuous climb toward greater truth; there are modes, normal science progresses smoothly, in crisis there is broad competition and then one wins and a field converges to normal progress again[Kuhn, 2012]. In a ”normal” science mode, scientist solve puzzles. The paradigm defines clear questions of mutual interest and value to the commu-

nity. Well defined questions lend to well developed, accepted experimental designs. However, there are other times when sciences can be in ”crisis”, this happens when the paradigm itself is under question. Competing paradigms arise. In this mode, scientists cannot rely on well developed experimental protocols and analyses to prove a result; providing evidence in support of a candidate paradigm is a categorically different task than filling in the blanks within an accepted one. Time of crisis mode science is when technology and science interface most directly; technological advances can reveal a crisis and others can reveal the solution. In this thesis, I aim to develop tools to allow psychology to be able to overcome current challenges. I test ideas in synthetic toy data and begin to evaluate in fMRI data. We will focus the contribution to three specific technical barriers that we need solutions for to advance in this area.

Chapter 2

Stability in Machine Learning

Often, in science, reanalysis of old datasets is an early step toward a new idea. Practically, some understanding of what is expected is required before resources for data collection are secured. Further, pilot studies on limited sample sizes in order to refine an experimental protocoal is also commonplace and a step toward good science. In both of these settings, it is important to glean something meaningful, if a lower burden of evidence out of a limited sample. Not enough data is a decreasingly common problem in many applications machine learning, but a very real constraint in science, where data is expensive. A common feature of these settings is that the quality of result necessary is different than others and what is most important is a prediction of how reproducible a result is. As a field of science, data analysis for psychology experiments is not just about the prediction portion of a machine learning problem, the interpretability of the model and is also important. Thus it is fitting to have performance measures that extend beyond classical losses applied to predicted values. In Machine Learning literature, stability has been used previously for generalization error bounds and a model selection criterion in applications, like clustering where loss functions are not obvious. We summarize this literature, identify the missing theory in order to bridge the gap and provide solutions for a number of reasonable cases that provide the analytical justification for the empirical, previously heuristic uses of stability measures. To do this we introduce a general form of a stability measure and illustrate that this structure induces a taxonomy over the various models that allows for easier comparison among prior results as well. Further we consider a theoretical result that appears to be inconsistent with empirical risk and determine the conditions under which it holds.

In particular, in psychology, a part of what has driven the need for a paradigm shift is the realization that experiments are far under-powered; b increasing the number of stimuli in an

experiment, it becomes evident that more of the brain is engaged in a simple attention task than previously believed and that that variety of engagement is also increased[Gonzalez-Castillo et al., 2012]. While many of the results that outline the theory presented in subsection 1.1.1 were due to changes in experimental design, the most striking change was to increase the sample size of the experiments. This means that in order to test other facets of the theory, we need to be conscious of the fact that we are using data that is too small and assess it appropriately. A performance measure that is sensitive to over-fitting is of great value.

Stability has been used in machine learning in a variety of capacities and with an even larger number of different definitions. In this chapter we collect the various definitions under a single framework for comparing the theoretical contributions and provide practical suggestions for leveraging this body of theoretical work in applications.

The objective of this chapter is to develop a principled presentation of stability as a framework for evaluating performance in machine learning applications. Stability is an appealing framework for real world problems because, using the framework we propose specific stability indices can be easily defined such that they are unit-interpretable in the domain and represent the priorities of the work. Learning theory results have demonstrated the theoretical utility of a variety of specific forms of stability and empirically stability has been used with success based on an intuitive justification. Here we provide a general form definition and the criteria under which complementary theoretical results hold for any instantiation. In a practical sense a stability analysis is a simple extension of existing analyses in many problems as it is an extension of cross validation or bootstrapping.

A number of contexts in which interpretable models are important is a growing area of machine learning. The impact of an algorithmic solution or other machine learning result is often dependent on the degree to which a human trusts the solution. In high risk application settings, people are not willing to hand off decision making to a machine if they are not able to understand what it's doing [Ridgeway, 2013](via [Zeng et al., 2015]) . This also applies to understanding the performance of a model. A measure that is both theoretically well-supported and intuitively valuable is a useful tool for both the machine learning researcher working jointly with scientists in other domains, and the practitioner using out of the box machine learning methods on their own data.

First, we provide a general framework for stability definitions and their properties(section 2.1). Then, we summarize previous work, using this common definition as a scaffold to compare how different definitions are used in deriving bounds, for model selection and for feature selection. Next, we present a principled view of how to use stability for model selection, through a probabilistic

interpretation of prior results and by filling necessary gaps to bridge from theory to practice.

In developing a general form of a stability definition we strive to define an interpretable framework that eases comparison among the numerous definitions in the literature. This will allow for each of the others to be recovered from the proposed template by a specific setting of the various components in the template definition. There are two main groups of stability definitions we aim to generalize: risk based definitions used for bounding generalization error and stability indices used for comparing feature selection results.

Most works provide a specific stability definition, or a number of specific definitions, but not a general form. [Giancarlo and Utro, 2012], Aims to provide an algorithmic paradigm and thus introduces a lot of generalization and structure. Working toward better clustering method assessment they pose that any stability analysis comprises four ingredients:

1. a data generation/perturbation procedure (sub-sampling/bootstrapping, noise injection, randomized dimensionality reduction, null hypothesis)
2. a similarity measure between partitions
3. a statistic on cluster stability
4. rules on how to select

We take a similar approach, but generalize to consider a broader class of learning algorithms and both asymptotic and empirical uses.

2.1 A Generic Stability Definition

To create a formal structure we propose that being *stable* is a theoretical or analytical property that an algorithm may have. Deriving from the nature of what it means to be stable, that for small changes in the input we can guarantee small changes at the output we propose that empirically, we can use an *empirical stability index* as a goodness of fit measure for a (data, algorithm) pair to be used for comparing models and assessing the match between the assumed model of an algorithm and the underlying phenomenon producing the data. A stability index, however, only applies as a a valid measure for a stable algorithm. This however, is not a restrictive property as numerous algorithms have been proven to be stable.

To consider the cases presented for use with generalization error, we first propose a general form of stability with respect to a loss function.

Definition 2.1. A *stability index* is a *dispersion* measure over a set of modified training sets. For training sets $\{S^{(1)}, S^{(2)}, \ldots, S^M\}$ of a sample $S \sim D^m$ and an algorithm $\mathcal{A}$:

$$I(A, S) = disp(\{S^{(1)}, S^{(2)}, \ldots, S^M\}) \tag{2.1}$$

Note that there is no specific relationship between m the sample size and M the number of modified training sets.

Definition 2.2. A learning algorithm, $\mathcal{A}$ is *stable* if a stability index converges to zero as the size of the training set increases.

Definition 2.3. The *stability* of a solution is the empirical estimate of a stability index for a given training set

Then, we can create a taxonomy of types of stability by choosing dispersion measures, convergences, and means of modifying the training set. This is just the most general form of how the variety of literature defines stability. This allows us to create three clear ways to compare the different notions of stability though: on their dispersion measure, type of convergence required, and type of training set modification. Stability indices do not depend on a type of convergence because they are a measure.

2.1.1 Dispersion Measures

A dispersion measure captures, the spread of the samples generated by applying the algorithm to the perturbed datasets. That is, for a perturbed dataset of size n the dispersion measure must map from $\mathcal{F}^n - \rightarrow \mathbb{R}$. This typically uses a pairwise distance among the resampled solutions or a distance of each them from some reference point, these are either averaged or the max is taken. These differences may be defined in the space of the function, the space of the parameters or with respect to the data sample.

A dispersion measure consists of a distance and a statistic. How they're defined and applied can vary. For example, range based dispersion measures take a difference (distance, since the minuend is known to be larger than the subtrahend) of two statistics, max and min, or third quartile and first quartile, etc. If the distance is applied first it may be pairwise among the samples or a distance from some reference point for each sample. The variance is the expectation of the distance from the mean of each point: taking euclidean distance and expectation as a statistic.

Definition 2.4. For models $\{f_i\}, i \in 1, \ldots, M$ a range dispersion measure:

$$d(T_1(\{f_i\}), T_2(\{f_i\})) \tag{2.2}$$

where d is any distance measure and T_1, T_2 are order statistic

Definition 2.5. For models $\{f_i\}, i \in 1, \ldots, M$ a pairwise dispersion measure:

$$T_{(i,j)\in[1,\ldots,M]}(d(f_i, f_j)) \tag{2.3}$$

where d is any distance measure and T is a statistic of the sample distances for every (i, j) pair

Definition 2.6. For models $\{f_i\}, i \in 1, \ldots, M$ and a reference model f_r a centered dispersion measure:

$$T(d(f_i, f_r)) \tag{2.4}$$

where d is any distance measure and T is a statistic of the sample-distances for each

Different notions of stability can be defined by their dispersion measure: a combination of a statistic and a distance. We will focus on expectation (arithmetic mean) and maximum or $|| \cdot ||_{infty}$ as statistics and the case where the statistic is applied to the distances. The distances are generated from the resampled solutions- either pairwise or relative to some notion of centrality. The training error is often used as a reference point to which all of the samples are compared. The distance between two pairs of solutions, or the solution and reference, may be defined in many ways. For example it can be a distance among functions directly, or any other representation. For example we may define distances in their parameter space or with respect to the sample through a loss function. If the learning algorithm $\mathcal{A}$ is an estimator, that is it represents elements of $\mathcal{F}$ using a set of parameters- it may be best to use a distance directly among the solutions. IF the learning algorithm produces a cluster labeling or decision rule, a more appropriate distance may be defined with respect to the samples.

Pairwise methods compute a quantity related to the diameter of the distribution, but even in leave-one-out cross validation, each pair is two samples different- thus introducing more variation in the input. A training fit centered method considers the radius of the 1-sample different solutions.

2.1.2 Training set Modification

This is typically a cross validation method, it also includes, for some prior literature, replace one re-sampling. Bootstrapping is also admissible. We will consider most closely k-fold

cross validation, where the data is split into k partitions, or folds, trained on $k-1$ and tested on the last part, and repeated to get a test score for each of the folds. We will focus especially the case where, for m samples, $k = m$ this is also called leave one out. Leave one out is especially appealing because it reduces to simple to express differences when using a training error centered dispersion measure. Asymptotically, there is little tradeoff by making this choice. For empirical assessment, classical challenges of how to partition a dataset of a fixed size for training and evaluation apply.

In the generalization literature, replace one is also sometimes used.

2.1.3 Convergence

Choosing a convergence type is useful for proving bounds, but not for creating a stability index for empirical use. We make some notes about different types of convergence to help relate the various forms of definitions found in the literature. First, convergence can be absolute or with various probabilistic forms and it can be in the form of an expectation or of a high probability. As when [Mukherjee et al., 2006] shows that their cross validation stability is equivalent, up to change of constants, to hypothesis stability go between an expectation form and a high probability form of the expression. Some authors do not include the convergence aspect directly in their definition, only that the quantity is bounded for a given sample size.

Review of contributions to learning that use concepts of stability using the unified framework to be able to compare different notions of stability and a probabilistic interpretation of the results- thus extending previous results from a probabilistic.

2.2 Generalization Error Bounds

In frequentist learning, the idea of a generalization error bound is appealing. This serves as a guide of how much to trust performance on a test set. Generalization error is the difference between the error on a test set and the *true* error, which can not be measured. In a Bayesian interpretation of probability, this is not even the objective, however, we will re-interpret some of these results as a means of deriving performance measures. Various notions of stability that are easy to show for a given class of algorithms have been used to bound generalization error in learning theory literature. Most of this work was done in the framework of empirical risk minimization.

Each of these results follows approximately the same storyline, introducing different stability indices to meet varying algorithmic needs.

1. Assume that for an algorithm A, $\Pr[I(A,S) \leq \beta] \geq 1-\delta$.
2. Derive $\alpha(\beta,\delta), \tau(\beta,\delta)$ such that $\Pr[|R - R_{emp}| \geq \tau] \leq \alpha$.
3. Show that for some algorithms of interest, $\beta(m) < \infty$.

These proofs do not rely on on knowledge of the distribution from which S is drawn. Additionally, β, α and τ will also depend on the sample size, m.

Where $R_{emp}(A,S)$ is the expected loss with respect to the sample and $R(A,S)$ is the expected loss with respect to the distribution. Given this general template a variety of different bounds using different stability indices are derived. The more strict the stability, the tighter bounds are derived. Generalization error bounds are derived based onthe method of bounded differences provided by [McDiarmid, 1989]. In [Kutin and Niyogi, 2002] a number of different bounds are derived for various strengths of stability, where these definitions correspond to various relaxations of the inequality below. By requiring it to hold strongly or weakly instead of uniformly, they are then able to prove bounds at three levels: uniformly, strongly and weakly and with respect to different quantities.

Theorem 1. *[McDiarmid, 1989] Let $X_1, \ldots, X_m$ be m i.i.d. random variables in a set S and assume that $F : S^m \to \mathbb{R}$ satisfies for all $i \leq m$:*

$$\sup |F(x_1, \ldots, x_m) - F(y_1, \ldots, y_{i-1}, y_i', y_{y+1}, \ldots, y_m)| \leq c_i \tag{2.5}$$

then

$$P(|F(Y_1, \ldots, Y_n) - E[F(Y_1, \ldots, Y_n)]| \geq \varepsilon) \leq 2e^{\frac{e^2}{\sum_{i=1}^m c_i^2}} \tag{2.6}$$

Uniform stability where $I(A,S) = ||\ell(A_S, z) - \ell(A_{S^{\setminus i}}, z)||_\infty$ provides a tight, simple bound.

Assuming an empirical risk minimization algorithm, these results require no assumptions about the distribution of the data. It requires that the cost function minimized is well behaved (σ-admissible) and uses how tight that behavior is to find the stability index bound, β. In addition to the algorithm having a bounded stability index, these results add additional constraints- to get constants in the bound.

The bounds in [Bousquet and Elisseeff, 2002] all require that the loss function, ℓ is bounded. Some express the boundedness in terms of the cost, were $\ell(f,z) = c(f(x),y)$ and require that $0 \leq c(y,y') \leq M$ for all y, y'. However in theorem 12 of that paper, the main result, the

bound is expressed as $0 \leq \ell(A_S, z) \leq M$ for all z, S. For all is a hard constraint and possibly unrealistic requirement to have on a problem, however, if we reconsider this for a realizable case, and that $\ell = -\log(p(z;\theta))$ then, this boundedness requirement can be interpreted as all of the samples having a small negative log likelihood, or high likelihood. This may not be guaranteed, but as an assumption for learning, the idea that under your distribution, none of the samples are extreme outliers is reasonable. This quantity M appears in the bounds- thus the quality of the solution is limited by the quality of the data.

A variety of algorithms have stability results proving that the generalization error is bounded. A typical result is in the form of assume that a stability index has a finite value, prove that makes the error bounded, and then show that for a class of algorithms the given index is in fact finite for a class of algorithms. Then show Uniform stability- a very strict version- has been shown for empirical risk of a bounded and σ-admissible loss function regularized in a RKHS.

2.3 Feature Selection

In feature selection, stability has been used to compare algorithms and stabilizing terms have been added to algorithms to improve task performance performance. Taking the large number of results on this topic together, we suggest that stability should not be used as a criterion for making general recommendations on the relative quality of algorithms, but rather as a metric for choosing which result to use in subsequent analyses after applying multiple algorithms to a dataset. That is, empirical stability measures agreement between the data and the model underlying the algorithm.

A popular stability index for feature selection is provided in [Kuncheva, 2007], rewritten to be consistent with our definition that index is as follows. It was initially defined using a similarity measure, with 1 for a stable solution. This measure is also normalized, so it does not depend on the training set size.

Definition 2.7. Let the output of a feature selection algorithm be a binary vector of length of the total number of features, d, $s\mathcal{A}_S \in \mathcal{B}^d$ where a 1 is in each entry for a retaine = d feature. Then for a set of solutions determined by n perturbed datasets $\{\mathcal{A}_{S^1}, \ldots, \mathcal{A}_{S^n}\}$ such that $|\mathcal{A}_{S^1}|_0 = \cdots = |\mathcal{A}_{S^n}|_0 = k$ where $0 < k < d$, using $|\cdot|_0$ as the zero "norm", the count of nonzero elements the Kuncheva [Kuncheva, 2007] consistency and stability indices are:

$$d(\mathcal{A}_{S^i}, \mathcal{A}_{S^j}) = 1 - \frac{|\mathcal{A}_{S^i}^T \mathcal{A}_{S^j}|_0 d - k^2}{k(d-k)} \tag{2.7}$$

$$I_{Kuncheva}(\mathcal{A}, S) = \frac{2}{n(n-1)} \sum_{i=1}^{n} \sum_{j=i+1}^{n} I_C(A_i, A_j) \tag{2.8}$$

In a comprehensive review of feature selection from an information theoretic point of view, [Brown et al., 2012] uses stability based on two different metrics ([Yu et al., 2008, Kuncheva, 2007] to compare algorithms. This work presents a number of information theoretic search criteria as approximations to maximizing the conditional likelihood of the training labels with respect to parameters θ indicating which features are on and τ for the distribution. Ultimately [Brown et al., 2012] presents average stability over a number of datasets to compare the search procedures, but in examining the per data set results, it shows different orderings of the different criteria. This phenomenon was also noted in [Kalousis et al., 2007], which conclude that no feature selection method is universally more stable than the others- they find that the ranking of the feature selection methods varies not by individual dataset, however, but by problem class- similar datasets provide similar rankings of algorithms on stability.

Other feature selection results on stability use a single dataset and use stability to choose the right feature selection criterion for that problem, or add stability into an algorithm.

2.4 Model Selection

Stability has also been used for model selection, especially in the case of unsupervised learning. The idea of stability using a risk function for model selection is developed in [Lange et al., 2002]. Using a sample-wise $0-1$ loss as a distance between solutions, and expectation as the statistic, this is developed as an empirical measure. This results in a metric being the expectation of the the number of disagreements when the model is trained on opposite halves of the data. Generalizing from 0/1 loss relative to the true answer, differences between models trained on separate halves of the data, this metric is finally extended to what is considered the most important case, unsupervised learning and examined for its ability to distinguish the right number of clusters. Toward the objective of a statistical theory of clustering, [Von Luxburg et al., 2005], suggests stability as a quantity of interest in clustering as generalization error is to classification.

Stability has been posed for both general model selection and model order selection[Lange et al., 2002]. In the case of determining the appropriate number of clusters, many examples of this

proving to be a reasonable metric working exist and intuitively, it makes sense. However, stability is not a sufficient criterion for the problem, Ben-David et al. [2006] show that for asymmetric clusters, incorrect numbers of clusters can appear stable, despite this method having been shown useful empirically.

In model selection for unsupervised learning, the similarity of solutions is defined as as similarity in the partitioning of the data, instead of through a sample-wise loss.

Given prior work using stability in machine learning, we ask the question, *how can we leverage the analytical results to provide motivation for the empirical uses?* Further, some of the above results are for restricted specific cases that vary per result- how can a unified definition aid in understanding these results more broadly. Stability as a general type of measure provides the opportunity to take this concept and define a distance such that the measure is interpretable in context. After we examine some general form results, we summarize advice on choosing a stability metric that gives the right type of insight.

In this section, we explore how to extend the existing results to be of more practical utility by using the general structure from section 2.1 to highlight the main differences To best extend these results, we look at how stability indices intended for empirical application are different from those for generalization error. There are two key differences: the type of distance measure (with respect to a samples versus in the parameter space) and the re-sampling method (replace or change one to use the method of bounded differences or a more varied re-sampling to produce maximum difference). We consider these difference and the gaps in necessary theory to overcome them in turn. The final step is to determine how an empirical estimate of a stability index varies from the analytical forms often defined.

For some (algorithm, stability index) pairs, the upper bound, β is easily derived. For example, if ℓ is a convex loss function with $0 \leq \ell() \leq M$, $|\ell(f_1, z) - \ell(f_2, z)| \leq \sigma |f_1(x) - f_2(x)|$, and $k(x, x) \leq \kappa^2$, for uniform stability $\beta \leq \frac{\sigma^2 \kappa^2}{2\lambda m}$ and $|| \cdot ||_k$ is a norm in a RKHS and the algorithm is[Bousquet and Elisseeff, 2002]:

$$A_S = \arg\min_g \frac{1}{m} \sum_{i=1}^{m} \ell(g, z_i) + \lambda ||g||_k^2 \tag{2.9}$$

To extend this to feature selection, we need to consider what the objective is there, in general. If considering candidate feature selection algorithms' results on a given dataset, it becomes a model selection problem. Following the problem setup where feature selection is an approximate conditional likelihood maximization, we aim to choose the algorithm that better approximates the

unknown distribution p. Then we will find the appropriate bound and finally examine for what algorithms we can extend this bound.

We do so by focusing on a probabilistic view of machine learning, taking the loss function as a negative log likelihood. This allows us to have a defined 'loss' even for unsupervised learning where a typical cost function is not well defined. By considering the case of MLE or MAP instead of a general (possibly regularized) empirical risk minimization we can see that the work in generalization bounds provides us a bound on the cross entropy between the true distribution q and candidate p. For example, taking one result of [Bousquet and Elisseeff, 2002], for uniform stability,

$$R < R_{loo} + \beta + \dots \tag{2.10}$$

$$E_Q[\ell(A_S, z)] \leq E_S[\ell] + \beta + \dots \tag{2.11}$$

$$\int -\log p(z;\theta) q(z) dz \leq -\sum_i \log p(z;\theta) + \beta + \dots \tag{2.12}$$

k

Taking the loss to be the NLL, we see that stability and the empirical likelihood, together, bound the cross entropy. Further, we can show that a stability is a necessary condition of having a KL divergence of zero. This shows that the cross entropy, which has the same minimizer as the KL divergence, is bounded by the likelihood of the data and the stability of the model. The constant M in [Bousquet and Elisseeff, 2002] is defined as $0 \leq \ell(A_S, z) \leq M$ for all samples, z, with ℓ as the negative log likelihood, this number is like the worse likelihood of any given sample. This number is smaller, and the bound is tighter if there are no outliers in the data. To bound this quantity, just means that all samples have nonzero likelihood under the model- which is a very reasonable assumption. Proving this bound for MAP, for the uniform stabiltiy in [Bousquet and Elisseeff, 2002] requires a convex likelihood and that for any two possible solutions θ_1, θ_2 and all possible samples z we have $|p(z, \theta_1) - p(z; \theta_2)| \leq \sigma|\theta_1 - \theta_2|$.

2.4.1 Distance Measures

In feature selection, the indices measure distance between solutions directly in a parameter space, not through a loss function. We propose that stability in the parameters implies stability in the likelihood, that is for a given form of distribution, a small change in the value of the parameter results in a small change of the likelihood of any sample in the domain.

A learning algorithm as defined is a mapping from a set of samples to some hypothesis. This can be represented in a number of different ways, each inducing a different distance measure. A flexible way of measuring disagreement, even when the hypotheses may be very different is to measure their distance relative to a sample- through a loss function. However, for a given learning algorithm, the different hypotheses produced by training on different often share a parametric form, and thus comparing them directly in the parameter space is possible.

If this holds for a specific distribution q then for an algorithm, A that is either an MLE or MAP of q, the the parameter stability of A on a sample $S \sim p$ is a guide for the cross entropy of the $H(p, q)$.

This is easy to check for any given distribution, as an example, take any member of the exponential family. Assume $|\theta_1 - \theta_2| \leq \beta_\theta$ and without loss of generality we can say that $\theta_1 = \theta_2 + \beta_\theta$. Then

$$p(z;\theta) = h(x) \exp \eta(\theta) \cdot T(x) - A(\theta)n \tag{2.13}$$

$$|p(z;\theta_1) - p(z;\theta_2)| = h(x)(\exp \eta(\theta_2 + \beta_\theta) \cdot T(x) - A(\theta_2 + \beta_\theta) - \exp \eta(\theta_2) \cdot T(x) - A(\theta_2)) \tag{2.14}$$

To understand this, we now just need to consider the form of $\eta(\theta)$ and $A(\theta)$ these are both linear combinations of $\theta^2, \log \theta$, and $\Gamma(\theta)$. If we understand how each of these functions behave when perturbed we get the behavior of the whole likelihood behaves.

$$(\theta + \delta)^2 = \tag{2.15}$$

2.4.2 Resampling

The second way in which the feature selection definitions are different is in resampling. They recommend using non-overlapping split-half sets instead of a training fit or leave one out. Therefore, we will need extensions of some of the inequalities that are used in the literature on generalization error bounds. All use McDiarmid's inequality to get the bound and to show that algorithms work, a requirement of σ admissibility or something similar is common as well. First we extend these ideas as needed for the case of our general definition and then we examine how these work.

In practice, choice of cross validation type is influenced by the size of the dataset, a less correlated estimate can be derived from fewer, larger folds instead of a leave one out. In this case, defining a stability notion with respect to different re-sampling methods may be advantageous.

Results that define the dispersion for the solutions due to re-sample via a loss function over the data can then bound directly the generalization error, which is also a function of the loss. In feature selection we actually want to consider a more strict case, not only that the estimated risk is close to the true risk, but that the feature list produced from the sample is similar to what would be produced on another sample or that it only varies with certain probability. As the empirical work in feature selection has shown, empirically, stability indices reveal properties of the data, even though a bound for these quantities can often be derived for classes of algorithms as well. Though a stability bound applies without regard for the generating distribution, the value

The results using stability to bound generalization error, show that the leave one out or empirical risk can be a good estimate of the expected value of the loss if taken with respect to the unknown distribution. The most general result showing a class of algorithms has a small $\beta(m)$ is from [Bousquet and Elisseeff, 2002]. It provides a bound for RKHS-norm regularized empricial risk minimization, with only the constraint that $0 \leq \ell(A, z) \leq M$ for all z and that $|\ell(f_1, z) - \ell(f_2, z)| \leq \sigma|f_1(x) - f_2(x)|$.

2.5 Connecting parameter space to sample space

We have from the analytical results that convergence of a stability index defined through a sample-space based distance and leave one out resampling shows that leave one out error is a bounded estimate true error. If we take negative log likelihood as a loss function then these results show that a stable solution means that the cross entropy estimate is good. Low cross entropy signals

Taking the case of loss functions that are negative log likelihoods as a starting point, we will show cases where bounded differences in parameter space yields bounded differences in the sample space.

Further, this shows that the stability of the solution bounds the likelihood of the sample under the true distribution and thus makes stability a goodness of fit measure. We can compute negative log likelihood in models without data, target pairs and so this means that stability can be used in unsupervised learning, up to corrections for permutation, where accuracy does not apply. This also overcomes the lack of units in negative log likelihood as the stability definition can be

First we show that if we take negative log likelihood as the loss, risk is cross entropy and empirical risk is its estimate.

First we consider the case of Gaussian with known, equal covariances. The Negative Log Likelihood (NLL) of each is the Mahalanobis distance of a test point to the mean. The Mahalanobis distance of these two means bounds the difference if the difference in Mahalanobis distance of any point to each of the two means, by the triangle inequality. So, if the means are close in Mahalanobis distance the two agree for all data points. With distributions centered at θ_1 and θ_2:

$$\ell_1(z) = -\log \mathcal{N}(\theta_1, \Sigma) \tag{2.16}$$

$$= \frac{1}{2}(z-\theta_1)^T\Sigma^{-1}(z-\theta_1) \tag{2.17}$$

$$= d_M(\theta_1, z)\ell_2(z) \qquad = d_M(\theta_2, z) \tag{2.18}$$

The triangle inequality gives:

$$|\ell_1(z) - \ell_2(z)| = |d_M(\theta_1, z)d_M(\theta_2, z)| \leq d_M(\theta_1, \theta_1) \tag{2.19}$$

So if $d_M(\theta_1, \theta_2) \leq \epsilon$, for any z, we have $|\ell_1(z) - \ell_2(z)| < \epsilon$. We do not need to find a bound in the mahalnobis distance for the means, if we have that for every dimension i, $|\theta_1 - \theta_2| < \delta$, then $\epsilon = \delta^2$.

2.6 Generalizing Resampling Methods

The bounds in Bousquet and Elisseeff [2002] use two inequalities to set the bounds. Both rely on sets of i.i.d. data points $S = \{z_1, z_2, \ldots, z_m\}$ and $S^i = \{z_1, z_2, \ldots, z_i^{'}, \ldots z_m\}$ and measurable functions $F : \mathcal{Z}^m \to \mathbb{R}$. These sets S and S^i are used to prove bounds relating to replace one resampled data sets. However, to be more robust, use of more general resampling is preferred by the heuristic applications in feature selection and clustering. Therefore, we aim to show extensions of these bounds for k-fold re-sampling of the data.

McDiarmid's inequality is used with the risk, $R(A, S) = E_z[\ell(A_S, z)]$ as $F(S)$.

In the resampling used for k-fold cross validation, the data is split into k partitions and the data is trained on each group of $k-1$ of those partitions. We can consider this as generating training sets that each leave out $n = \frac{m}{k}$ samples. For the results to work on these folds, we need to

redefine the necessary theorem in terms of S^{n_i} where n samples are replaced. To do this, we will introduce a random variable T which is a function of a set S. Since R_e is the sample mean of the loss, we can compute it over batches of smaples and then average those together. R depends on teh expectation with respect to z. If the algorithm A can operate in batches.

We need to define a combinations function that takes a group of samples $x_{(i-1)n+1}, \ldots, x_{ni}$ and aggregates them into a single random variable $z_i = f(x_{(i-1)n+1}, \ldots, x_{ni})$. The requirements of f are that z_i are iid. Assuming that the data x_i are iid, then these z_i are also iid for any linear function f. If A_S is the same when computed from z or from x then these cases are the same and using k-fold is equivalent, up to a change in the bound to be shown below. Any algorithm that can be computed in batches and accumulated by a linear function meets this requirement.

Chapter 3

Bayesian Nonparametric Methods for Structure Discovery in Neuroimaging Analysis

As noted, neuroimaging to date has been plagued by the hard implicit assumptions made by the analysis techniques available in popular neuroimaging analysis applications. In this chapter, we propose Bayesian nonparametric techniques that extend these methods and extend and customize existing Bayesian nonparametric methods to uniquely address challenges of this data.

Bayesian nonparametric models are a form of probabilistic reasoning that allows for more flexible classes of models. A parametric statistical model is one with an a priori fixed number of parameters, everything else is nonparametric. In Bayesian probability, this means that the *number* of parameters is not fixed. This allows the model complexity to adapt to the complexity of the observed data. As a simple model, if we knew that brains only did ten different things and any other signal variation was noise, we could apply a standard clustering algorithm to the data to find the signature of each of those 10 concepts and assign all of our trials to one of those ten concepts. However, in reality, we do not know how many thing the brain can do, and further, we have little hope that in a given experiment we observe everything or even how many we observe in that experiment. Instead, a Bayesian nonparametric model assigns trials to groups based on similarity, but can add new groups or merge samples it initially thought was two groups if it begins to seem more likely that there is one group. This provides an even more important advantage over time, if more data is collected later, we can use what we learned before about the concepts and assign new points to those, but we

probably also observed some new concepts so we can add those as well.

We summarize our objective here as structure discovery- we will use largely unsupervised methods. First we consider that we do not know the structure of the experiment completely a priori- while of course we know enough of the observed things, because they were set, since we do not know how subjects interpret stimuli, we can relax that assumption. We can discover hat structure that relates stimuli or trials to each other. Next we relax the assumption that we know the response shape- we can discover the structure of a stimulus response as well. Finally we consider that the time-course is more unknown, not only that a single trial response is an unknown, but that the relationship among responses is unknown.

3.1 Standard fMRI Analysis General Linear Model

The generalized linear model is the most standard model for analyzing functional MRI data. This model carries a lot of strong assumptions, however, many of these are also no longer fully validated in the present. in this chapter we develop further relaxations and generalizations of this classical model.

The GLM is linear in that the main model is that for experimental design, X and measurements, y, we can learn regressors β by solving

Neural Model, for N_j stimuli of condition j at times $\tau_i, i = 1, \ldots, N_j$

$$s_j(t) = \sum_i^N \delta(t - \tau_i)$$

Response model, for a known hemodynamic response, $h(t)$:

$$a_j(t) = s_j(t) * h(t)$$

Design matrix for C conditions, nuisance signals M, and fmri acquisitions at times κ

$$A = [a_1(\kappa), \ldots, a_C, M, \mathbf{1}]$$

Data model

$$A\beta = y$$

This assumes that we can construct a good model for the experiment in X and that all regions of the brain (represented by voxels here) follow the same response. The most common model is that there is a spike of neural electrical activity in important regions and that then those

regions have an increased need for blood. This is modeled as a delta at the stimulus onset (t_s) convolved with an a priori selected, spatially uniform HRF, $h(t)$ that is sampled at the acquisition times. Sometimes the increase in electrical activity is modeled with a step function, indicating a sustained increase in activity for an amount of time, but this is still the convolved with the HRF.

$$X_i(t) = h(t) * \delta(t_s) \tag{3.1}$$

1. That psychological concepts are localized to small portions of the brain
2. That there is a spatially uniform BOLD response (neural convolved with hemodynamic response- we cnnot separate these readily)
3. That the experimenter assigned stimulus labels best capture similar neural responses variability.
4. That trials are independent.

This first idea we test largely through experimental design- by increasing the experimental power, we hypothesize a larger percentage of the brain to have a significant correlation with task.

3.2 Related Methods in Neuroimaging

Clustering has been applied in neuroimaging before in many context. Specifically, clustering has been applied in studies where unconstrained hemodynamic response shapes are used in data analysis as a post processing step Neta et al. [2015], Gonzalez-Castillo et al. [2012]. Our proposed model instead learns the hrfs as members of clusters. Prior work learns a single mean hrf for each voxel independently- with noise varying from trial to trial and then learns a cluster mean in a secondary step. Instead, our method says that in each voxel, for each trial, is an hrf for the 'region' that the voxel belongs to.

Other work has relaxed the assumption of known hemodynamic responses, most often through a finite impulse response function. This works well for a dataset with time-locked measurements, but without, there is a lot of averaging in time. Using the GP for the time model allows us to model a function smooth in time with a coarse sampling, as is done in standard models using a fixed HRF, while still being flexible through the uncertainty model.

Mixture models for GLMs in fmri have been consideredPenny and Friston [2003].

3.3 Related Models in Alternate Contexts

There are many works that mix nonparametric clustering priors with Gaussian processes. First we review the literature related to these preliminary components, then the similar models.

3.3.1 Dirichlet Process Mixture Model

The most classical nonparametric model is the Chinese restaurant process. This is the colloquial name for a posterior distribution over the cluster assignments when the prior is a Dirichlet process. This can be conceptualized as an infinite extension of the most common clustering model: the Gaussian Mixture Model. A clustering model for data x and assignments z has two main components: a likelihood of a sample, given the cluster assignment ($P(x_i|z_i)$) and a prior on the cluster assignment ($P(z)$). There is also often a prior on the cluster shape and location parameters- the parameters of $P(x_i|z_i)$, a mean and covariance for Gaussian clusters. However here we are more interested in the structure of a prior for cluster assignments. If we knew there were two clusters, a distribution over them could be a Bernoulli distribution: $P(z = 1) = \rho, P(z = 0) = 1 - \rho$. If we do not know the relative probability of the clusters a priori, we put a prior on ρ A Bernoulli distribution has two parameters that control its shape. These hyper parameters of our model control our rough idea about what the relative size of the cluster, but in a probabilistic way, We can set these parameters so that we are likely to get a larger cluster and a smaller cluster, but we do not have to make a hard decision.

For more than two clusters we can generalize from a Bernoulli to Multinoulli, or categorical, distribution for $P(z)$ and we can generalize the Beta to Dirichlet distribution for $P(\rho)$. A sample drawn from a Dirichlet distribution is a vector of length K with non-negative elements that sum to one, which is exactly the parameter we need for the Multinoulli distribution to sample cluster assignments from. We can imagine that $K \to \infty$ and then we have what is called the Dirichlet Process. This would allow us to hypothesize an infinite number of possible latent components, but of course when we sample assignments from that for N data points, we would only observe some finite number, $K^+ \leq N$, of clusters. In order to draw samples, we only actually need the distribution $P(z)$. The CRP is the distribution for this. It's name comes from the following analogy. Imagine a Chinese Restaurant with an infinite number of tables. Each customer enters and chooses a table. Customer $i = 0$ has to start a new table. The next customer ($i = 1$) chooses the same table as customer 1 with probability $\frac{1}{1+\alpha}$ or a new table with probability $\frac{\alpha}{1+\alpha}$. Each subsequent customer chooses a table with probability proportional to the number of people at that table and a

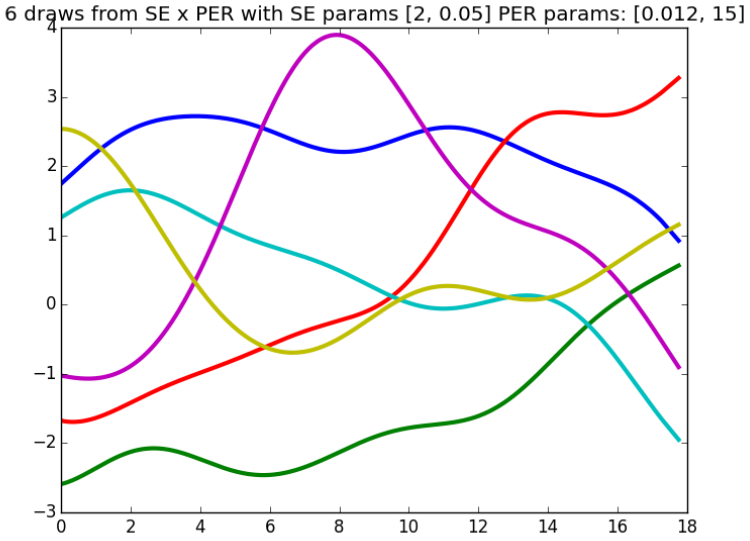


Figure 3.1: Sample draws from a GP with the same kernel that is used for data analysis

new table with probability proportional to α that is customer i chooses a table with $c < i$ people with probability $\frac{c}{i+alpha}$ and starts a new table with probability $\frac{\alpha}{i+alpha}$.

3.3.2 Gaussian Processes

Gaussian processes provide a flexible model for specifying a prior over a functionRasmussen and Williams [2006]. That means a draw from a Gaussian process is a function, as showing in Figure 3.1 Imposing a structured, additive kernel allows for an unsupervised method for decomposing additive and temporal structure in data[Duvenaud et al., 2013]. This model proposes that signals can be decomposed into additive and multiplicative combinations of basic functions- then uses a Gaussian process prior and additive and multiplicative structure in the kernel to model this behavior[Duvenaud et al., 2011]. A similar, but more structured approach is in [Fox and Dunson, 2012] where a tree structure to model a multi-scale additive GP prior is used to describe Magnetoencephlogram (MEG) signals. This is used toward the objective of detecting changepoints- where changepoints are modeled like frequency changes. We use an multiplicative Kernel to provide ample structure to relate to the data that we have observed and control it through easy hyper parameters.

3.3.3 Gaussian Process Clustering

Gaussian Processes have been clustered with a Dirichlet Process prior by Hensman et al. [2015], however this model is for more structured data, using a sum and mean for different clusters. Our model instead clusters GPs in one dimension of the data and assumes that each time series is a variably weighted repetition of the same GP. In our work, the main role of the GP is to provide a latent continuous model for combining information among variably sampled trials. Though the sampling rate is continuous, the trial onsets are not time-locked to this sampling.

Gaussian Processes have been used as an observation model in clustering posed as a data association problemRoss et al. [2014], Ross and Dy [2013], Schulam. Specifically or disease trajectory sub-typing, Assuming a number of i.i.d input, output pairs drawn from a collection of latent functions. The GP is used to describe the function and thereby cluster the input out put pairs. Our model, however we consider the samples to be a noise observation of a latent function, not a single point. We already know which points are associated in the time direction, but there are many time series and we want to cluster them using the GP as the mean.

3.3.4 Multiple clustering

Other works have also put multiple clustering priors on data, most often these are hierarchichalTeh et al. [2006b], Cadez and Smyth [1999]. The Dual Beta Process Prior modelRoss et al. [2014] allows for multiple memberships rather than clusters but has similar assignment priors over the two directions of hte data matrix. Ours differs however in that we allow only a single assignment and that our proposed mean is a product of two means for a time window instead of a single output based on the GP assigned latent function.

3.4 Learning Trial similarities, spatially variable hrf

In machine learning a popular class of methods for discovering latent structure in data is Bayesian nonparametric models. These are statistical models that derive from Bayesian view of probability, so all unknown quantities are considered to be random variables (latent variables and parameters) and probability is interpreted as uncertainties. A parametric statistical model is one that describes a system or dataset that has a *finite* number of parameters, fixed a priori and independent of the size of the data. A nonparametric model is one where the number of parameters of the model is not finite, in this setting, the number of parameters will grow with the size of the data. Note

that nonparametric models, in general, can also be distribution free or have non finite number of parameters in varying ways. Bayesian nonparametrics are a specific type of nonparametric model and differ from ()

Our proposed model for fmri data is similar to the standard model, but instead of being constrained to a fixed model, we will learn more of the structure from the data. Typically, areas that are engaged in representing the stimulus are assumed to have an increase neural activity in a region, and the BOLD signal is be modeled as a hemodynamic response convolved with the firing, by averaging trials deemed similar by the experimenter the common activation is assumed to be due to the common property of those trials.

Our model has a varied level of amplification on the hrf with the type of trial. However, we are not dictating a priori which trials we expect to have similar underlying spatial activation patterns, we will learn this from the data. Further, we do not assume a fixed hemodynamic response across the entire brain. We assume that, relative to the number of voxels the whole brain has a small number of different HRFs.

3.4.1 Assumptions

1. the hrf varies spatially, a given region (voxel) responds with the same hrf for all stimuli though
2. a 'condition' corresponds to a weighting of the hrfs spatially throughout the brain, which ones do/do not respond or relative weights

3.4.2 Notation

- $n \in [1, \ldots, N]$ trial index, $N \leq 900$ depending on partition of data used
- $t_n \in \mathbb{R}^{a_i}$ sample acquisition times within the duration of trial i,
- $a_i \in [4, 6]$ number of fMRI acquisitions in each trial, trial duration varies 14-18s (4s stimulus and 10-14 uniformly random fixation) and TR=3 (sampling rate)
- $v \in [1, \ldots, V]$ voxel index, for V voxels
- $z_i \in [1, \ldots, Z]$ trial condition for trial i and Z observed conditions
- α_z concentration parameter for trial clustering (learned conditions), in N trials, expected number of conditions is $\mathbb{E}[Z]\alpha_z \log N = 6$ (picked based on stimulus labels in valence/arousal)

- $u_k \in [0, 1]$ stick breaking weights for z
- $\pi_k \in [0, \ldots, 1], \sum_{k=1}^{\infty} \pi_k = 1$ mixing proportions for trials, z
- $\beta_z \in \mathbb{R}^V$ is the spatial activation map for condition z
- $\mu_0 \in \mathbb{R}$ mean 'activation' level of a voxel $\mu = 0$
- $\sigma_0 \in \mathbb{R}$, variance for voxel weights σ, broad
- $r_v \in [1, \ldots, D]$ region for voxel v and D regions across the whole brain
- α_r concentration parameter for region discovery, in V voxels, expected number of regions is $\mathbb{E}[D]\alpha_r \log V = 20$ (picked to match G-C best FIR clustering)
- $w_j \in [0, 1]$ stick breaking weights for z
- $\rho_j \in [0, \ldots, 1], \sum_{k=1}^{\infty} \rho_j = 1$ mixing proportions for trials, z
- $h_r(t)$ is the hrf for region r, a continuously valued function
- K is kernel function for GP, likely set to SE $\times$ PER with parameters for visual smoothness of hrfs, shared across all regions (for now)
- $m(t) \in \mathbb{R}^6$ (known) motion parameters at time t
- $\mu_{t,v} \in \mathbb{R}$ the voxel specific mean for run including time t
- $\eta_v \in \mathbb{R}^6$ learned motion weights per voxel and run?
- ϵ is noise variance , set to something small

3.4.3 Generative Process

$$u_k \sim \text{Beta}\,(1, \alpha_u) \tag{3.2}$$

$$\pi_k = u_k \prod_{j}^{k-1}(1 - u_j) \tag{3.3}$$

$$z_i \sim \text{Mult}\,(\pi) \tag{3.4}$$

$$\beta_{z_i,v} \sim \mathcal{N}(\mu_0, \sigma_0^2) \tag{3.5}$$

$$w_j \sim \text{Beta}\,(1, \alpha_w) \tag{3.6}$$

$$\rho_j = w_j \prod_{l}^{j-1}(1 - w_j) \tag{3.7}$$

$$r_v \sim \text{Mult}\,(\rho) \tag{3.8}$$

$$h_r(t) \sim GP(0, K) \tag{3.9}$$

$$y_v(t_i) \sim \mathcal{N}(h_{r_v}(t_i)\beta_{z_i,v} + \eta_v m(t_i) + \mu_{j,v}, \epsilon) \tag{3.10}$$

We can conceptualize this as learning the design matrix, relative to a standard GLM model. Instead of pre-specifying which trials, we will learn where to place the hrfs in different conditions. Also, instead of using a single design matrix for the entire brain, we are using different design matrices for different voxels. Not a unique one per voxel, but learning which groups of voxels shoud share.

3.5 Inference

Standard GLM procedures compute a maximum likelihood solution, a point estimate of the unknown that makes the data likelihood as high as possible. We instead want to find a postertior distribution of the latent variables give then osberved data. We can take a point estimate from that distribution for post analysis, but we want to compute the full posterior distribution, $p(\beta, r, h, z|\mathcal{D})$. Using the generative process and Bayes rule we get:

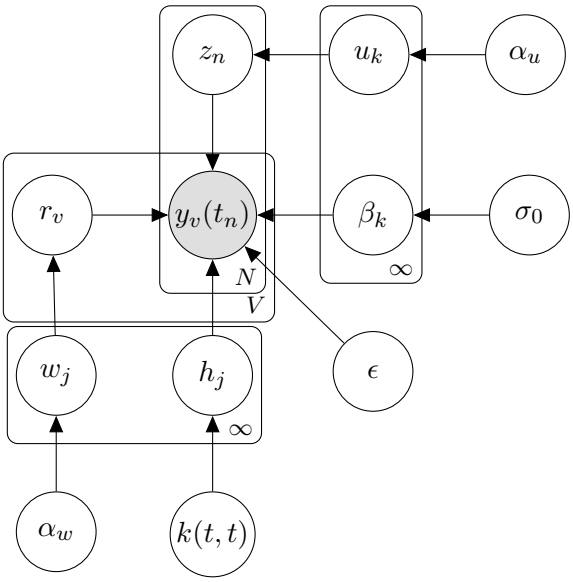


Figure 3.2: Probabilistic graphical model representation of the proposed generative model.

$$p(\beta, r, h, z|\mathcal{D}) = \frac{p(\mathcal{D}|\beta, r, h, z)p(\beta|z)p(z|u)p(h|r)p(r|w)}{p(x)} \tag{3.11}$$

$$\propto \prod_{v}^{V}\prod_{i}^{N} p(y_v(t_i)|\beta_{v,z_i}, h_{r_v}, \epsilon)p(\beta_{z_i,v}|z_i, \mu_0, \sigma_0)p(z_i|u)p(u|\alpha_u)p(h_{r_v}|r_v, K)p(r_v|w)p(w|\alpha_w) \tag{3.12}$$

The denominator can be written as an integral of the numerator, but it is intractable, so we must use an approximate inference technique. There are two main approximation techniques: variational inference and sampling. We choose variational inference because it is deterministic as opposed to sampling to save computational time.

Variational inference changes from an inference step that is computationally intractable, generally due to an infeasibly expensive or algebraically intractable normalization constant into a *variational* optimization, meaning an optimization over functions. To simplify, we pick a class a functions q as our *variational distribution* and optimize its parameters to match the desired distribution, minimizing the KL divergence. This procedure learns distribution, not a point estimate of parameters because it optimizes the parameters of another distribution to match the target. We will consider our model where all of the latent variables are included in θ so that $\theta = [\mathbf{Z}, \beta, \mathbf{R}, \mathbf{h}]$ and

Variational inference maximizes the Evidence Lower Bound (ELBO):

$$\log p(y) = \log \int_{\theta} p(y, \theta) \tag{3.13}$$

$$= \log \int_{\theta} p(y, \theta) \frac{q(\theta)}{q(\theta)} \tag{3.14}$$

$$= \log \left(\mathbb{E} \left[\frac{p(y, \theta)}{q(\theta)} \right] \right) \tag{3.15}$$

$$\geq \mathbb{E} \left[\log(p(y, \theta)] - \mathbb{E} \left[[] \log q(\theta) \right] \right. \tag{3.16}$$

Now we choose the partition of the hidden variables.To do this, we must choose an approximating distribution, this is most commonly done by assuming partitions of the hidden variables, $\theta = [\mathbf{Z}, \beta, \mathbf{R}, \mathbf{h}]$ (both latent variables in the model and the parameters) are independent and assuming distributions for those.

$$q(\theta|\nu) = \prod_{m}^{M} q_m(\theta_m|\nu_m) \tag{3.17}$$

We will use the mean field approximation and assume each θ_m is a single variable. where θ_{-m} means all elements of θ except for i. We choose to partition the variables so that each of the original variables gets its own q_m and corresponding variational parameter, ν^{θ_m}. Then our approximation is as follows, with truncations in the variational distribution at level T (for the trials) and S (for the regions):

$$q(\beta, r, h, z|\mathcal{D}) = \prod_{t=1}^{T-1} q_{\tilde{u}_t}(u_t) \prod_{t=1}^{T} q_{\tilde{\beta}_t}(\beta_t) \prod_{i=1}^{D} q_{\tilde{z}}(z_i) \prod_{s=1}^{S-1} q_{\tilde{w}_s}(w_s) \prod_{s=1}^{S} q_{\tilde{h}_s}(h_s) \prod_{i=v}^{V} q_{\tilde{r}}(r_v) \tag{3.18}$$

The variational solution is then the update the following expectation

$$\ln q_m^*(\theta_m) = \mathbb{E}_{l \neq m}[\ln p(X, \theta)] + c \tag{3.19}$$

In the exponential family, we can use standard forms to compute each step. We will find that, the optimal distributions for the above are $q_{\tilde{u}_t}(u_t)$ and $q_{\tilde{w}_s}(w_s)$ are Beta distributions, $q_{\tilde{z}}(z_i)$ and $q_{\tilde{r}}(r_v)$ are multinomial distributions, $q_{\tilde{\beta}_t}(\beta_t)$ are Gaussian and $q_{\tilde{h}_s}(h_s)$ are Gaussian processes. The variational free parameters are:

$$\nu = [\tilde{u}_1, \ldots, \tilde{u}_T, \tilde{\beta}_1, \ldots, \tilde{\beta}_T, \tilde{z}_1, \ldots, varz_D, \tilde{w}_1, \ldots, \tilde{w}_S, \tilde{h}_1, \ldots, \tilde{h}_S, \tilde{r}_1, \ldots, \tilde{r}_V] \tag{3.20}$$

Distributions in the exponential family include Dirichlet, Gaussian, Beta, and Categorical; all of the distributions used in our model.

We use coordinate ascent to maximize the ELBO, where in each variational parameter one at a time, we update the

The updates for $\tilde{u}$ and $\tilde{w}$ are the standard DP stick breaking variational updatesBlei and Jordan [2004].

$$\tilde{u}_{t,1} = 1 + \sum_{i}^{D} \tilde{z}_{i,t} \tag{3.21}$$

$$\tilde{u}_{t,2} = \alpha_u + \sum_{i}^{D} \sum_{j=t+1}^{T} \tilde{z}_{i,j} \tag{3.22}$$

$$\tilde{w}_{s,1} = 1 + \sum_{v}^{V} \tilde{r}_{v,t} \tag{3.23}$$

$$\tilde{w}_{s,2} = \alpha_w + \sum_{v}^{V} \sum_{j=s+1}^{S} \tilde{r}_{v,j} \tag{3.24}$$

$$\tag{3.25}$$

To compute the updates we need to compute trial-wise estimates of h and voxel-wise estimates of β, marginalized across regions and conditions respectively, we denote these by with hats.

$$\hat{h}_{\mu,v} = \sum_{s} \tilde{r}_{v,s} \tilde{h}_{s,\mu} \tag{3.26}$$

$$\hat{h}_{\Sigma,v} = \sum_{s} \tilde{r}_{v,s} \tilde{h}_{s,\Sigma} \tag{3.27}$$

$$\hat{\beta}_{\mu,n} = \sum_{t} \tilde{z}_{n,t} \tilde{\beta}_{t,\mu} \tag{3.28}$$

$$\hat{\beta}_{\Sigma,n} = \sum_{t} \tilde{z}_{n,t} \tilde{\beta}_{t,\Sigma} \tag{3.29}$$

the updates for $\tilde{h}$ and $\tilde{\beta}$ are analogous to mixture of regression mean updates, in the mean variance parameterization.

$$\tilde{\beta}_{t,\mu} = \tilde{\beta}_{t,\Sigma} \sum_n \frac{\tilde{z}_{n,t}}{epsilon} \hat{h}^T_{\mu,v} y_n \tag{3.30}$$

$$\tilde{\beta}_{t,\Sigma} = \left(\sum_n \frac{1}{\epsilon} \operatorname{Tr}\left(\hat{h}_{\Sigma,v}\right) + \frac{1}{\epsilon} \hat{h}^T_{\mu,v} \hat{h}_{\mu,v} + \frac{1}{\sigma_0} I_V \right)^{-1} \tag{3.31}$$

$$\tilde{h}_{s,\mu} = \tilde{h}_{s,\Sigma} \sum_v \frac{\tilde{r}_{v,s}}{\epsilon} \hat{\beta}_{\mu,:,v} y_v \tag{3.32}$$

$$\tilde{h}_{s,\Sigma} = \left(\sum_v \frac{1}{\epsilon} \tilde{r}_{v,s} \hat{\beta}^T_{\mu,:,v} I_{N_a} \hat{\beta}_{\mu,:,v} + \frac{1}{\epsilon} \operatorname{Tr}\left(\hat{\beta}_{\Sigma,n}\right) + K^{-1} \right)^{-1} \tag{3.33}$$

The updates for $\tilde{r}$ and $\tilde{z}$ are appropriately modified distances.

$$\mathbb{E}\left[[] \log u_i\right] = \psi(\tilde{u}_{i,1}) - \psi(\tilde{u}_{i,1} + \tilde{w}_{t,2}) \tag{3.34}$$

$$\mathbb{E}\left[[] \log(1 - u_i)\right] = \psi(\tilde{u}_{i,2}) - \psi(\tilde{u}_{i,1} + \tilde{u}_{t,2}) \tag{3.35}$$

$$\tilde{z}_{n,t} \propto \exp\left(\mathbb{E}\left[\log u_t\right] + \sum_i^{t-1} \mathbb{E}\left[\log(1 - u_i)\right] + \sum_v \frac{1}{\epsilon} \beta_{t,v} \hat{h}^T_{v,\mu}(t_n) y_{v,n} - \frac{\beta^2_{t,v}}{2\epsilon} \hat{h}^T_{v,\mu}(t_n) \hat{h}_{v,\mu}(t_n) - \frac{\beta^2_{t,v}}{2} \operatorname{Tr}\left(\hat{h}_{\Sigma,v}(t_n)\right) \right) \tag{3.36}$$

$$\mathbb{E}\left[\log w_s\right] = \psi(\tilde{w}_{s,1}) - \psi(\tilde{w}_{s,1} + \tilde{w}_{s,2}) \tag{3.37}$$

$$\mathbb{E}\left[\log(1 - w_s)\right] = \psi(\tilde{w}_{s,2}) - \psi(\tilde{w}_{s,1} + \tilde{w}_{s,2}) \tag{3.38}$$

$$\tilde{r}_{v,s} \propto \exp\left(\mathbb{E}\left[\log w_s\right] + \sum_i^{s-1} \mathbb{E}\left[\log(1 - w_i)\right] + \sum_n \frac{1}{\epsilon} \hat{\beta}_{\mu,n,v} \tilde{h}^T_{s,\mu}(t_n) y_{v,n} - \frac{\hat{\beta}^2_{\mu,n,v}}{2\epsilon} \tilde{h}^T_{s,\mu}(t_n) \tilde{h}_{s,\mu}(t_n) \frac{\hat{\beta}^2_{\mu,n,v}}{2\epsilon} \operatorname{Tr}\left(\hat{h}_{\Sigma,v}(t_n)\right) \right) \tag{3.39}$$

We iterate the above until the the ELBO converges. The ELBO is:

$$\mathcal{L} = E_q[\log p(y, \beta, h, z, r, u, w)] + E_q[\log q(y, \beta, h, z, r, u, w)] \tag{3.40}$$

$$= \sum_{v,i} E_q \log p(y_v(t_i) | \beta_{z_i,v} h_{r_v}) \sum_{t,i} E_q \log p(z_i = k) + \sum_t E_q \log p(\beta_k) + \sum_t E_q \log p(u_t) \sum_{s,v} E_q \log p(r_v = \tag{3.41}$$

3.6 Model Checks

For this model, we do not have a *ground truth* that we are trying to recover. Instead we will use qualitative assessments to interpret the results. Further, we have explicitly chosen to not model some of the structure that we believe should be there, but had too much uncertainty about, these unmodeled items we can then use as assessments to check the model output against our hypotheses. Within each subject, we will fit portions of the data to the model. To assess the model there are a number of post-analysis steps that we can do, each based on different questions. For each solution we learn four main pieces of information a partition of the voxels by hrf shape, an hrf shape for each group of voxels, a partition of the trials, a spatial activation pattern for each trial type.

3.6.1 Does the inference algorithm work?

First we will assess the inference algorithm with synthetic data. The synthetic data are sampled from the generative model, with some checking to ensure diversity on the beta maps. In this case we can compare the recovered posteriors to the sampled variables and assess the likelihood of the ground truth under the posterior. Additionally, from this assessment we can compare initialization schemes and convergence.

3.6.2 HRFs and regions

We can analyze the learned Gaussian process means, the $\tilde{h}_\mu$ to see if these shapes are biologically plausible by visual inspection. We can also compare the learned shapes qualitatively to those of other studies that allowed for a varied hemodynamic response across the brainGonzalez-Castillo et al. [2012], Moriguchi et al. [2011]. We can also compare the learned responses from different subsets of a subjects data and across subjects. We do not expect them to exactly match, subjects can have differences and a single subject might have some variability from session to session especially since these are collected on different days, up to months apart. However, we expect there to be similarities. Further, we expect the response in some brain regions, to be similar to the canonical hrf that is often used in standard analyses.

We can also compare the actual brain parcellations and assess how certain these parcellations are. We check the entropy per voxel of these assignments to assess the certainty of the assignment. These maps can be compared to anatomical images. We expect some of these regions

to match anatomical regions and to be continuous in space and others to represent long spatial range functional correlation.

3.6.3 Trials and spatial activations

In the spatial activations a first question of interest is if the maps for any of the conditions partition similarly to the region maps based on hrfs. That is to say, do the regions, defined by temporal response shape activate as a whole per condition or more variably. As above we can assess stability across samples within a subject and across subjects. We can also assess entropy per trial, how many trials are certain, how many are similar. Further, we can use the many labels that exist including basic affect, semantic content and physical properties like hue to compute mutual information with the trial condition assignments learn from the brain activity in order to provide psychological interpretations of the conditions.

3.7 Results

We present results first challenging the idea that there are small isolated regions that respond ti images based on experimenter determined labels, by increasing the number of trials. This verifies that there is mismatch in the standard techniques. Then we present results of evaluating the proposed dual DP model on synthetic data for validation of inference code and characterization of sensitivity to initialization. Finally we show the limited performance on real data and assess further what changes to initialization and algorithm settings might be best based on the synthetic results.

3.7.1 General Linear Model

Experiments with the standard analysis results confirm that for affective tasks as well, the amount of general engagement with a simple model increases as more data is added as shown previously for attentional tasksGonzalez-Castillo et al. [2012]. For these results, we executed the following mass averaging procedure:

1. Concatenate n runs of data and design matrices for $n = 2, 7, 12, \ldots N$
2. Solve least squares
3. Compute t statistic and p value

4. Apply Bonferoni multiple comparisons correction

5. Count number of voxels with $p_B < .05$

The design matrices include the model, six motion parameter time courses (x,y,z each for translation and rotation) and a column of ones for the mean, a sample is shown in Figure 3.3. We completed this for models of novelty (all stimuli as one condition), valence (3 levels) and arousal(2 levels). For each level n the procedure is completed five times and the results are show in Figure 3.4-Figure 3.14. For the valence results the arousal levels are collapsed and for arousal the valence levels are collapsed.

3.7.2 Double DP Synthetic Results

For synthetic results we varied the initialization, the order of the updates and repeated for multiple random restarts for each version of the initialization and multiple random draws. Overall, solutions are better with respect to the activation maps($\tilde{\beta}_\mu$) and condition assignments ($\tilde{z}$) and more variable with respect to the hrfs ($\tilde{h}\mu$). The solution for the region assignments ($\tilde{r}$) is not as good. This, however could be partially attributed to the fact that these are not sampled from a prior that encourages diversity, often they are similar, and only the most distinct ones are recovered.

Synthetic data is sampled from the model with αs set so that the expected number of regions is 20 and the expected number of conditions is six. We use GPs draws sampled with a Kernel function of Squared Exponential with parameters 2 and .05 times Periodic with parameters .02 and 15. This was chosen to be visually reasonable given the data and prior knowlendge of hemodynamic models. Samples are shown in Figure 3.1. The β maps are drawn with a heuristic to enforce diversity instead of directly from the model prior. For each condition a ρ is drawn and then β_v are drawn from $\mathcal{N}(1, \sigma_0/5)$ and with probability ρ mulitplied by -1. This makes the majority distant from zero, though the overall is still zero mean and the negative beta weights make for a more diverse and distinct activation map.

The order of the updates and the initialization are related, varying the order of the updates, which initialization is more influential changes. We compare four different orders switching either updates over time or space first and within each the updates of assignments or cluster parameters, the update orders are listed in Table 3.1.

We tested each inference variation with each of four initialization options. Within these we applied initialization very close to the truth, slightly noisier in the assignments, close assign-

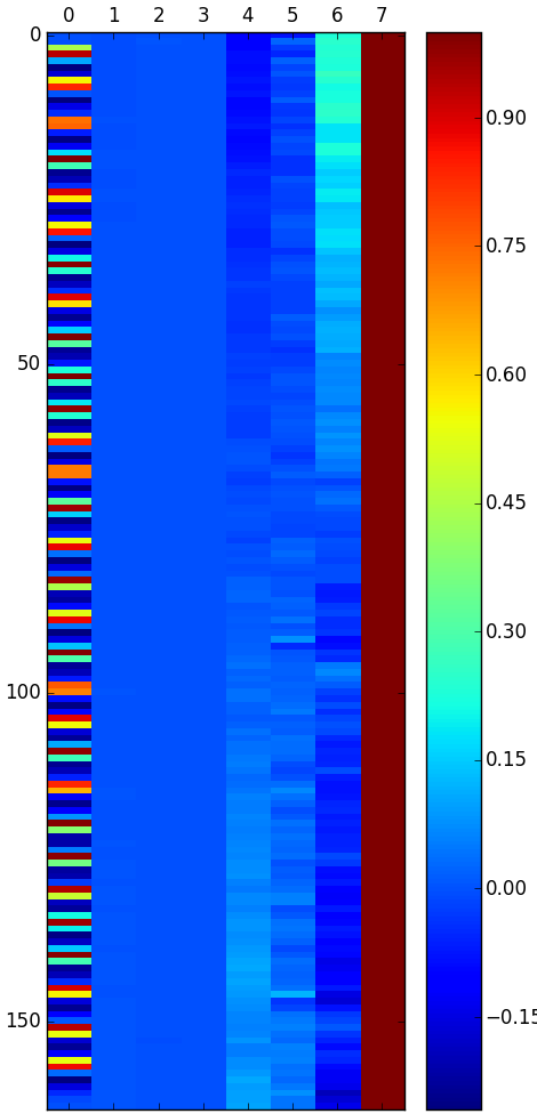


Figure 3.3: Sample design matrix for novelty

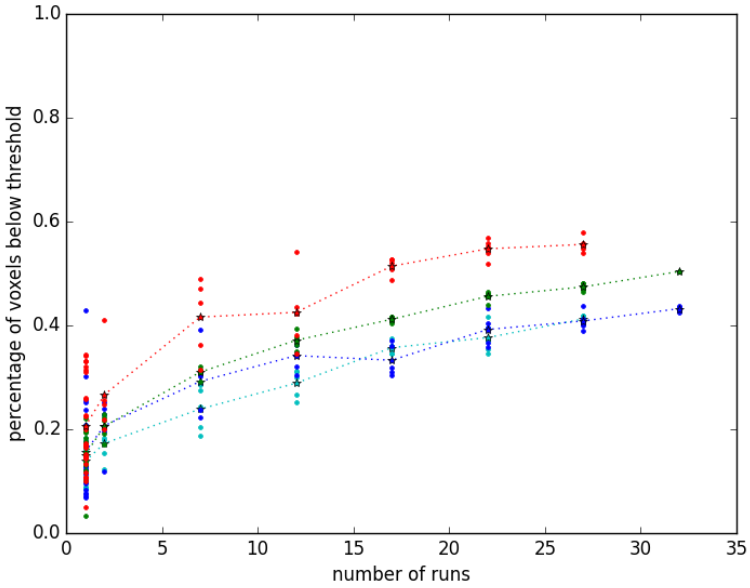


Figure 3.4: Results of Mass Averaging for novelty. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.

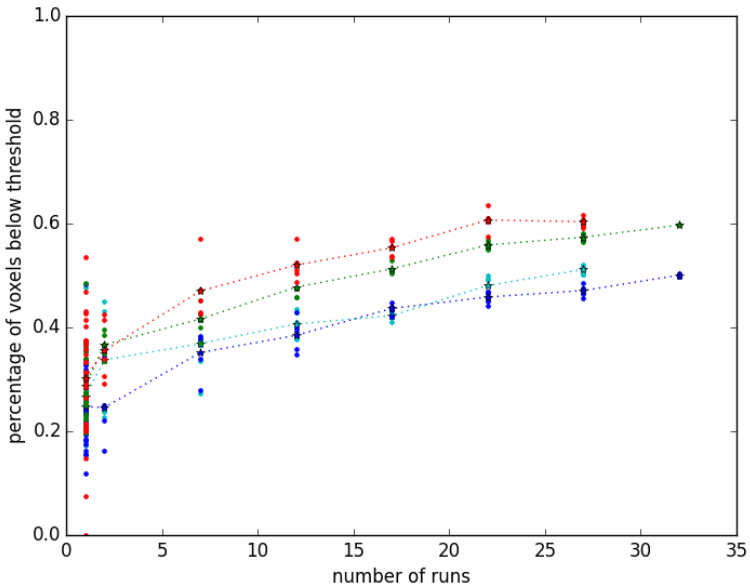


Figure 3.5: Results of Mass Averaging for high arousal. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.

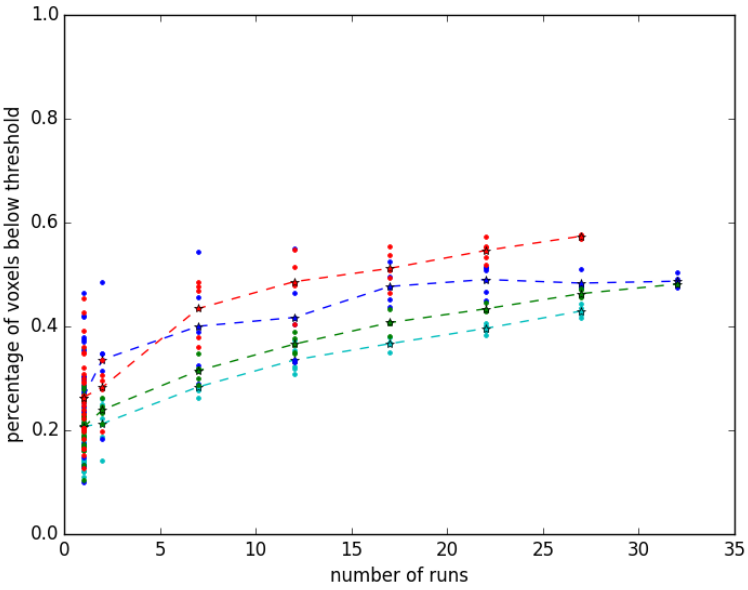


Figure 3.6: Results of Mass Averaging for low arousal. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.

update step	space,properties	space, assignments	time, properties	time assignments
1	$\tilde{h}$	$\tilde{w}$	$\tilde{\beta}$	$\tilde{u}$
2	$\tilde{w}$	$\tilde{r}$	$\tilde{u}$	$\tilde{z}$
3	$\tilde{r}$	$\tilde{h}$	$\tilde{z}$	$\tilde{\beta}$
4	$\tilde{\beta}$	$\tilde{u}$	$\tilde{h}$	$\tilde{w}$
5	$\tilde{u}$	$\tilde{z}$	$\tilde{w}$	$\tilde{r}$
6	$\tilde{z}$	$\tilde{\beta}$	$\tilde{r}$	$\tilde{h}$

Table 3.1: Listing of the update orders in the four inference variations

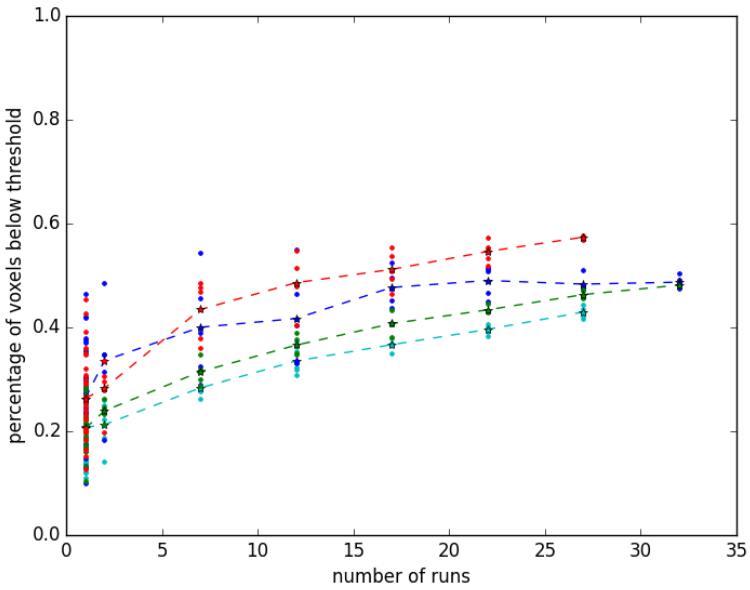


Figure 3.7: Results of Mass Averaging for the intersection of voxels active in both high and low arousal. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.

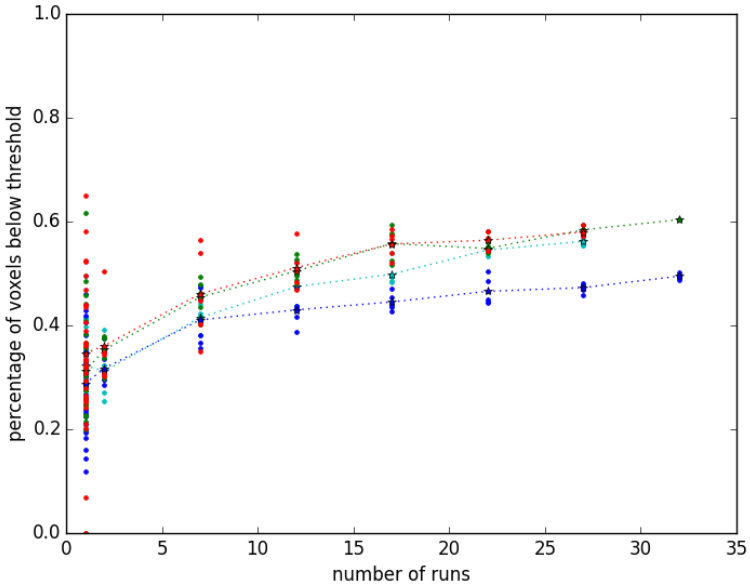


Figure 3.8: Results of Mass Averaging for negative valence. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.

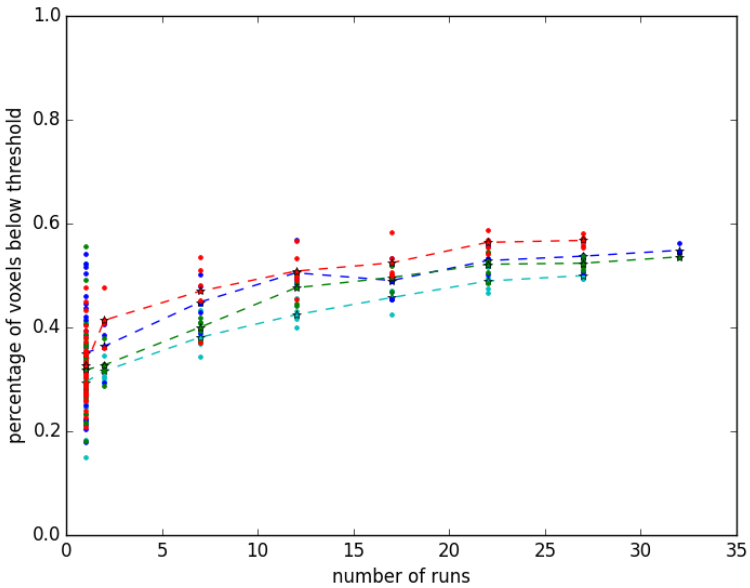


Figure 3.9: Results of Mass Averaging for neutral valence. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.

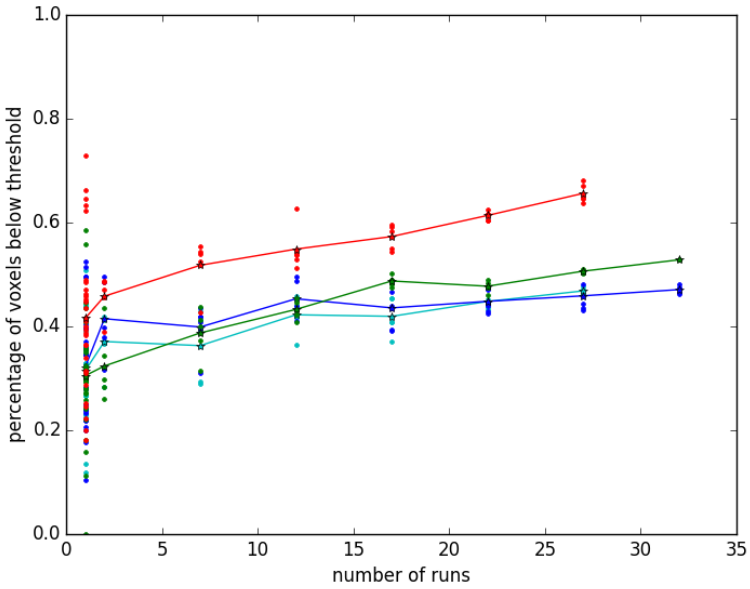


Figure 3.10: Results of Mass Averaging for positive valence. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.

ments with only probabilistic seeds, and noisy assignments with probabilistic seeds. The initialization includes settings for $\tilde{\beta}$, $\tilde{h}$, $\tilde{z}$ and $\tilde{r}$. In all cases we sample $\tilde{z}$ and $\tilde{r}$ from a Dirichlet distribution. We set the truncation level for each to 4 times the expected number of clusters based on the α_u and α_w, giving $T = 24$ conditions and $S = 80$ regions. Using the true r and z we sample from $\text{Dir}\,(a_0 + r_{0,s})$ and $\text{Dir}\,(a_0 + z_0)$ for each voxel and trial. The base concentration a_0 is the appropriate α as a vector and r_0 and z_0 are vectors zero everywhere except for a at the index r_t and z_t. In the noisy case the coefficients are adjusted to increase the entropy of each sample examples of these initializations appear in Figure **??** and Figure **??**.

In the lower noise cases, $\tilde{\beta}_\mu$ and $\tilde{h}_\mu$ are initialized to the truth with samples drawn from the respective priors appended. In the probabilistic only cases, the true means are not included, only draws from the prior are used, though the assignmetns are still derived from the truth. In all cases $\tilde{\beta}_\Sigma$ and $\tilde{h}_\Sigma$ are initialized to the outer products of their respective means. For the cases with minimal noise good solutions are achieved from all random restarts of the initialization for all inference cases. Updating in time or space seems to have minimal impact, but updating the cluster properties first if not well initialized has detrimental outcome.

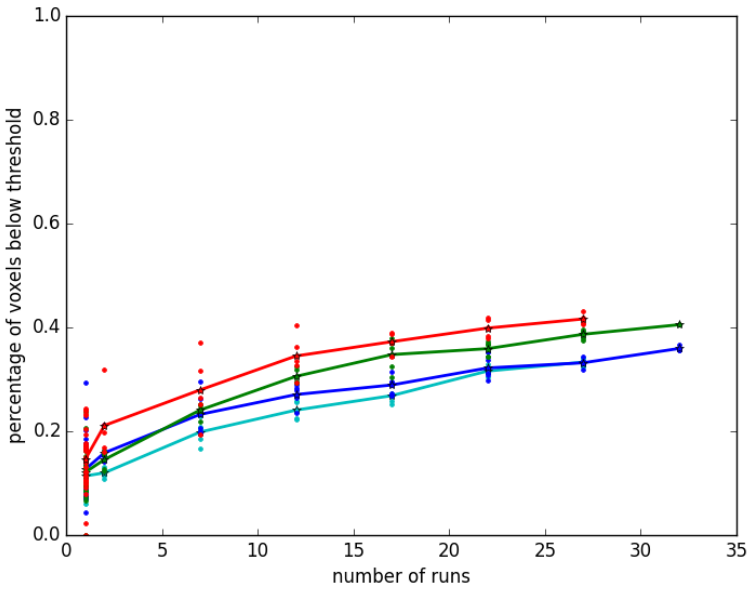


Figure 3.11: Results of Mass Averaging for the intersection of voxels active in negative and neutral valence. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.

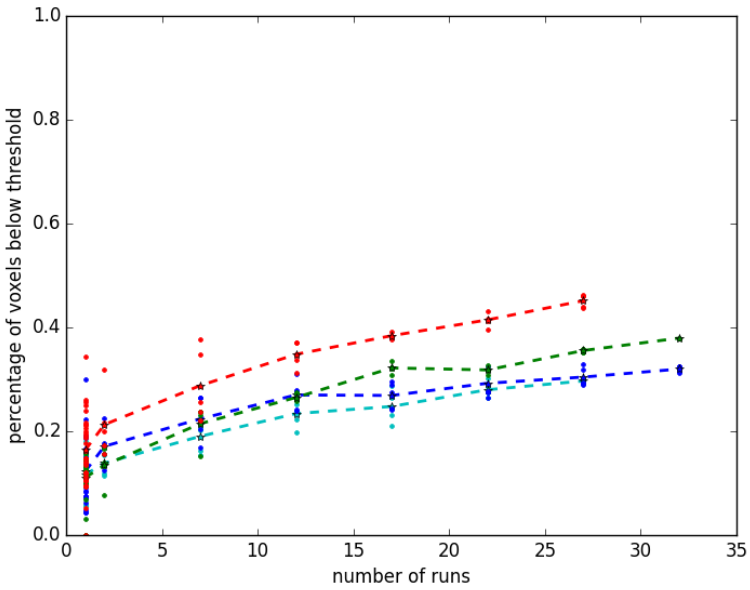


Figure 3.12: Results of Mass Averaging for the intersection of voxels active for negative and positive valence. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.

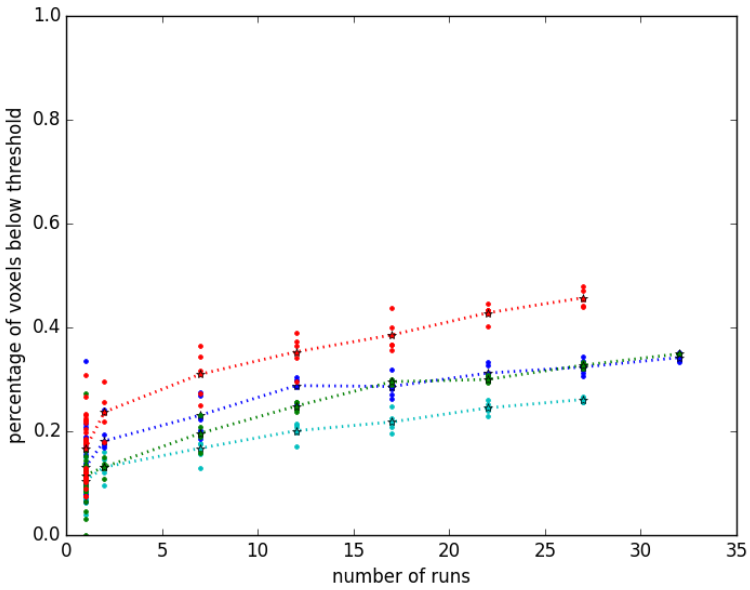


Figure 3.13: Results of Mass Averaging for the intersection of voxels active for neutral and positive valence. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.

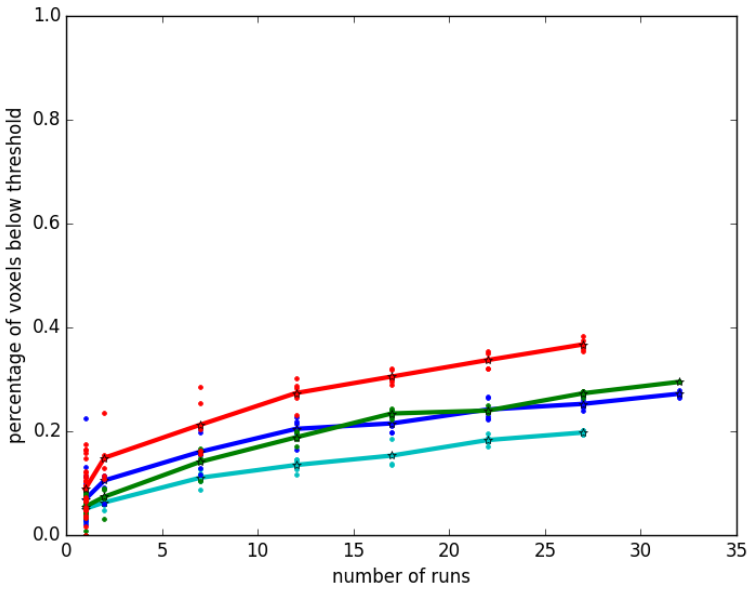
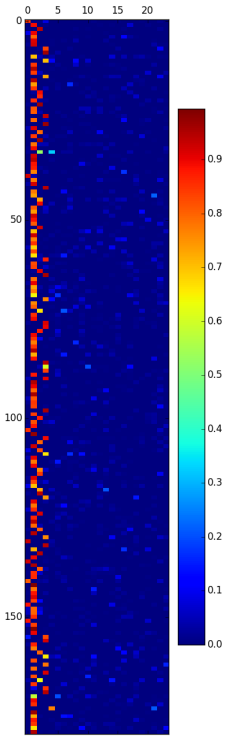
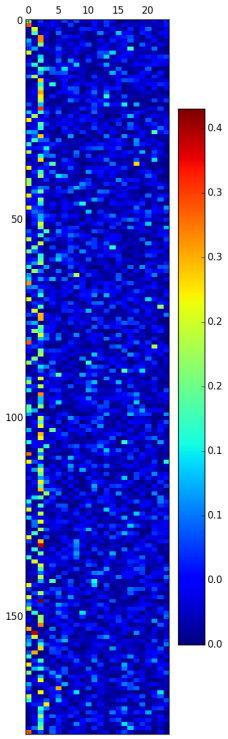


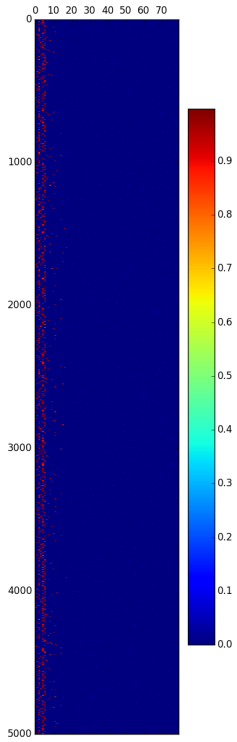
Figure 3.14: Results of Mass Averaging for the intersection of voxels active for positive, negative and neutral valence. Colors represent difference subjects, dots indicate an individual solutions and the trace shows the mean.



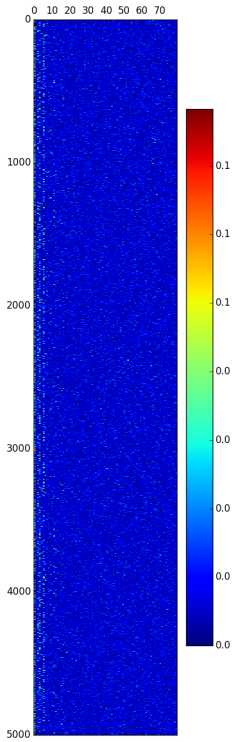
(a) Sample close to truth initialization for condition assignments



(b) Sample noisy initialization for condition assignments



(a) Sample close to truth initialization for region assignments



(b) Sample noisy initialization for region assignments

3.7.3 Double DP fMRI Results

In the real data, the inference is extremely sensitive to the initialization. Further work is needed to determine an appropriate initialization scheme. With the existing schemes and settings, after a single update of $\tilde{h}$ all of the functions are nearly zero everywhere, no structure is left. In the context of the more advanced synthetic analysis and varied initialization schemes, it seems that a prior for diversity and better region assignment seeds may help. Additionally, it is probably most robust to start update the time clustering model first, then the spatial model. The activation maps, β are much less sensitive to poor initialization.

The initialization for real data is a very noisy $\tilde{z}$ based on the stimuli labels, and moderately noisy $\tilde{r}$ based on the Harvard Oxford cortical atlas. The initializations for $\tilde{\beta}$ and $\tilde{h}$ are drawn from the prior. After 1 or 2 iterations each time the solution is stuck at a degenerate case, spatially. All voxels assigned to regions approximately the same way and the corresponding hrfs are nearly flat. Further exploration of the initialization space is left to future work. Based on synthetic results we will use the time first and update assignments using data prior to updating the cluster properties.

Chapter 4

Conclusion and Discussion

In conclusion, my thesis develops computational tools designed to address an new class of challenges in psychology research. By focusing on interpretable models either through simplicity or carefully generative explanations, we develop both context sensitive and generalizable methods for interpreting data from psychology experiments. After a collaborative effort to design an abstract mathematical model for a new theory of mind-brain mapping, I have developed computational techniques that allow for re-analysis of experiments designed in the current tradition for interpretation in the context of a new direction. Through modeling data, I will extend capabilities for analysis and interpretation of other complex systems, beyond the brain and by considering the real challenges to this transient mode of science operation, I will provide a framework for design of context-appropriate performance measures that can be re-used in a broad variety of applications.

Ideally this work serves as a case study in exploring how machine learning methodology changes when the objective is discovery in highly uncertain scientific contexts. Increasingly, machine learning researchers are adopting this objective and beginning to design methods for novel domains and assess them in context-sensitive ways. These efforts are not always as well appreciated[Wagstaff, 2012], but are gaining momentum, (increase of workshops). Matters that in a idealized data, analytical sense are just details, are actually complex problems and a real barrier to driving true breakthroughs. While current knowledge still forces these steps to be explored in individual contexts, it is unlikely that there is no transferable knowledge from for example, genetics to neuroscience.

There is opportunity for things that are not the most exciting fronts from the perspective of one domain to be the necessary impetus to be a catalyst to a breakthrough in another domain. in machine learning we are uniquely skilled to be able to influence a lot of different domains of

scientific and social understanding.

4.1 Real Data Results

This model does not yet run fully successfully on real data. With the completed synthetic analysis and continued conversations and exploratory methods we will seek an improved initialization. The initialization needs to be one such that the updates for $\tilde{h}_\mu$ do not reach a divergent solution. We will explore sampling the GP to initialize with a prior for diversity[Zou and Adams, 2012], as well as finer grained anatomical parcellations as the initialization for the region assignments.

Recent advances in variational inference have proposed adding steps to better accommodate the possibility of merging and splitting clusters[**?**]. Further in both sampling and variational setups collapsed algorithms where parameters are marginalized perform better[Teh et al., 2006a]. Either or both of these could improve over the specific challenges faced in the inference on the real data.

4.2 Future work in UFBMM framework

There are a number of additional directions in which we can relax traditional assumptions to approach the idealized model that that the UFBMM provides. Natural extensions of the dual DP model proposed here in light of the framework might be to extend from each voxel always having the same response shape, with varying amplitude to each voxel having a small set of possible shapes. This could be accomplished using a topic model prior instead of the clustering prior used now. The framework also suggests a fruitful choice may be to expand from a single clustering assignment to each trial to a feature allocation prior, where a given trial may have multiple labels, perhaps influencing different areas of the brain.

Bibliography

”Analysis Group at FMRIB”. FSL FLIRT FAQ. URL http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FLIRT/FAQ#Can_I_register_to_an_image_but_use_higher.2BAC8-lower_resolution_.28voxel_size.29.3F.

”Analysis Group at FMRIB”. Fslutils. URL http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Fslutils.

Lisa Feldman Barrett. *How emotions are made: The secret life the brain.* Houghton-Mifflin-Harcourt., New York, NY, 2017.

Lisa Feldman Barrett and Ajay Bhaskar Satpute. Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. *Current opinion in neurobiology*, 23(3):361–372, 2013.

Lisa Feldman Barrett, Dana Brooks, Sarah M Brown, Elizabeth Clark-Polner, Jaume Coll-Font, Jennifer G Dy, Deniz Erdogmus, Burak Erem, Ajay B Satpute, and Christine D. Wilson-Mendenhall. Inferring the Mind by Modelling the Brain: A Framework for Computational Modeling. *Manuscript in Progress*, 2017a.

Lisa Feldman Barrett, Dana Brooks, Sarah M Brown, Elizabeth Clark-Polner, Jaume Coll-Font, Jennifer G Dy, Deniz Erdogmus, Burak Erem, Ajay B Satpute, and Christine D. Wilson-Mendenhall. Inferring the Mind by Modelling the Brain : Beyond Faculty Psychology. *Manuscript in Progress*, 2017b.

Shai Ben-David, David Pal, and Ulrike v. Luxburg. A Sober Look on Clustering Stability. (2002), 2006. ISSN 03029743. doi: 10.1007/11776420_4. URL http://eprints.pascal-network.org/archive/00003907/.

David M. Blei and Michael I. Jordan. Variational methods for the Dirichlet process. *International Conference on Machine Learning*, page 12, 2004. ISSN 1936-0975. doi: 10.1145/1015330.1015439. URL http://portal.acm.org/citation.cfm?doid=1015330.1015439.

Olivier Bousquet and Andre Elisseeff. Stability and Generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002. ISSN 15324435. doi: 10.1162/153244302760200704. URL http://dl.acm.org/citation.cfm?id=944801$\delimiter"026E30F$nhttp://www.crossref.org/jmlr_DOI.htmlhttp://dl.acm.org/citation.cfm?id=944801.

Gavin Brown, Mikel Luj, A Pocock, MJ Zhao, M Luján, and Mikel Luj. Conditional likelihood maximisation: A unifying framework for information theoretic feature selection. *The Journal of Machine Learning Research*, 13:27–66, 2012. URL http://dl.acm.org/citation.cfm?id=2188387.

Igor Cadez and Padhraic Smyth. Probabilistic clustering using hierarchical models. (99), 1999.

Gustavo Deco, Giulio Tononi, Melanie Boly, and Morten L Kringelbach. Rethinking segregation and integration: contributions of whole-brain modelling. *Nature Reviews Neuroscience*, 2015.

David Duvenaud, H Nickisch, and CE Carl Edward Rasmussen. Additive Gaussian processes. *arXiv preprint arXiv:1112.4394*, pages 1–9, 2011. URL http://arxiv.org/abs/1112.4394.

David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Zoubin Ghahramani. Structure discovery in nonparametric regression through compositional kernel search. In *Proceedings of the International Conference on Machine Learning (ICML)*, volume 30, pages 1166–1174, 2013. URL http://arxiv.org/abs/1302.4922.

Emily B Fox and David B Dunson. Multiresolution Gaussian Processes. In F Pereira, C J C Burges, L Bottou, and K Q Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 737–745. Curran Associates, Inc., 2012. URL http://papers.nips.cc/paper/4682-multiresolution-gaussian-processes.pdf.

Raffaele Giancarlo and Filippo Utro. Stability-based model selection for high throughput genomic data: an algorithmic paradigm. In *Artificial Immune Systems*, pages 260–270. Springer, 2012.

Javier Gonzalez-Castillo, Ziad S Saad, Daniel a Handwerker, Souheil J Inati, Noah Brenowitz, and Peter a Bandettini. Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis. *Proceedings of the National Academy of Sciences of the United States of America*, 109(14):5487–5492, April 2012. ISSN 1091-6490. doi: 10.1073/pnas.1121049109. URL http://www.ncbi.nlm.nih.gov/pubmed/22431587.

Krzysztof Gorgolewski, Christopher D Burns, Cindee Madison, Dav Clark, Yaroslav O Halchenko, Michael L Waskom, and Satrajit S Ghosh. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in neuroinformatics*, 5(August):13, 2011. ISSN 1662-5196. doi: 10.3389/fninf.2011.00013. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3159964&tool=pmcentrez&rendertype=abstract.

Douglas N Greve and Bruce Fischl. Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, 48(1):63–72, 2009.

Sten Grillner, Nancy Ip, Christof Koch, Walter Koroshetz, Hideyuki Okano, Miri Polachek, and Mu-ming Poo. Worldwide initiatives to advance brain research. *Nature Publishing Group*, 19 (9):1118–1122, 2016. ISSN 1097-6256. doi: 10.1038/nn.4371. URL http://dx.doi.org/10.1038/nn.4371.

James Hensman, Magnus Rattray, and Neil D. Lawrence. Fast nonparametric clustering of structured time-series. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):383–393, 2015. ISSN 01628828. doi: 10.1109/TPAMI.2014.2318711.

Mark Jenkinson and Stephen Smith. A global optimisation method for robust affine registration of brain images. *Medical image analysis*, 5(2):143–156, 2001.

Mark Jenkinson, Peter Bannister, Michael Brady, and Stephen Smith. Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841, 2002.

Mark Jenkinson, Christian F Beckmann, Timothy E J Behrens, Mark W Woolrich, and Stephen M Smith. FSL. *Neuroimage*, 62:782–790, 2012. doi: 10.1016/j.neuroimage.2011.09.015.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007.

Thomas S Kuhn. *The structure of scientific revolutions*. University of Chicago press, 2012.

LI Kuncheva. A stability index for feature selection. *Artificial intelligence and applications*, pages 390–395, 2007. URL http://www.actapress.com/Abstract.aspx?paperId=29484.

S Kutin and P Niyogi. Almost-everywhere algorithmic stability and generalization error. *In UAI-2002: Uncertainty in Artificial Intelligence*, pages 275–282, 2002. URL http://dl.acm.org/citation.cfm?id=2073909.

PJ Lang, MM Bradley, and BN Cuthbert. International affective picture system (IAPS): Technical manual and affective ratings. Technical report, NIMH Center for the Study of Emotion and Attention, 1999.

Tilman Lange, Mikio L ML Braun, Volker Roth, and Joachim M Buhmann. Stability-based model selection. In *Advances in neural information processing systems*, pages 617–624, 2002. URL http://machinelearning.wustl.edu/mlpapers/paper_files/AA17.pdf.

Stacy Marsella, Jonathan Gratch, and Paolo Petta. Computational models of emotion. In *A Blueprint for Affective Computing-A Sourcebook and Manual*, pages 21–46. Oxford University Press, 2010. URL http://books.google.com/books?hl=en&lr=&id=C2gLOQ105okC&oi=fnd&pg=PA21&dq=Computational+Models+of+Emotion&ots=-KeLZ-bqyS&sig=3sRulbXWgqLemNgjjMj1qW341Oc.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

Yoshiya Moriguchi, Alyson Negreira, Mariann Weierich, Rebecca Dautoff, Bradford C Dickerson, Christopher I Wright, and Lisa Feldman Barrett. Differential hemodynamic response in affective circuitry with aging: an FMRI study of novelty, valence, and arousal. *Journal of Cognitive Neuroscience*, 23(5):1027–1041, May 2011. ISSN 1530-8898. doi: 10.1162/jocn.2010.21527. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3141584&tool=pmcentrez&rendertype=abstract.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1-3):161–193, July 2006. ISSN 1019-

7168. doi: 10.1007/s10444-004-7634-z. URL http://link.springer.com/10.1007/s10444-004-7634-z.

KP Murphy. *Dynamic bayesian networks: representation, inference and learning*. Phd, University of California, Berkeley, 2002. URL https://bitbucket.org/bmmalone/library/src/b5f06a50b629/Murphy2002.pdfhttp://onlinelibrary.wiley.com/doi/10.1002/cbdv.200490137/abstracthttp://higherintellect.info/texts/science_and_technology/artificial_intelligence/DynamicBayesianNetworksRepresenta.

Maital Neta, Francis M Miezin, Steven M Nelson, Joseph W Dubis, Nico U F Dosenbach, Bradley L Schlaggar, and Steven E Petersen. Spatial and temporal characteristics of error-related activity in the human brain. *The Journal of Neuroscience*, 35(1):253–266, 2015. ISSN 0270-6474.

Brian Patenaude, Stephen M Smith, David N Kennedy, and Mark Jenkinson. A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922, 2011.

Will Penny and Karl Friston. Mixtures of general linear models for functional neuroimaging. *IEEE Transactions on Medical Imaging*, 22(4):504–514, 2003. ISSN 02780062. doi: 10.1109/TMI.2003.809140.

Carl Edward Rasmussen and K.I. Williams. *Gaussian processes for machine learning*. 2006. ISBN 026218253X. doi: 10.1142/S0129065704001899.

Greg Ridgeway. The pitfalls of prediction. *NIJ Journal*, 271:34–40, 2013.

James C Ross, Michael H Cho, Jennifer G Dy, and Peter J Castaldi. Dual Beta Process Priors for Latent Cluster Discovery in Chronic Obstructive Pulmonary Disease. *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14*, pages 155–162, 2014. doi: 10.1145/2623330.2623750.

Jc Ross and Jg Dy. Nonparametric Mixture of Gaussian Processes with Constraints. *Jmlr.Org*, 28, 2013. URL http://jmlr.org/proceedings/papers/v28/ross13a.pdf.

Peter Schulam. A Framework for Individualizing Predictions of Disease Trajectories by Exploiting Multi-Resolution Structure. *Advances in Neural Information Processing Systems*, pages 1–9.

Stephen M Smith. Fast robust automated brain extraction. *Human brain mapping*, 17(3):143–155, 2002.

Stephen M Smith and J Michael Brady. SUSANa new approach to low level image processing. *International journal of computer vision*, 23(1):45–78, 1997.

Michael Spivey. *The continuity of mind*. Oxford University Press, 2008.

Yee W Teh, David Newman, and Max Welling. A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *Advances in neural information processing systems*, pages 1353–1360, 2006a.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 101(476):1566–1581, December 2006b. ISSN 0162-1459. doi: 10.1198/016214506000000302. URL http://www.tandfonline.com/doi/abs/10.1198/016214506000000302.

William R Uttal. *The new phrenology: The limits of localizing cognitive processes in the brain.* The MIT press, 2001.

Ulrike Von Luxburg, Shai Ben-david, and Ulrike Von Luxburg. Towards a statistical theory of clustering. In *Pascal workshop on statistics and optimization of clustering*, pages 20–26, 2005. doi: 10.1209/epl/i2004-10507-8. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.91.6237&rep=rep1&type=pdf.

Kiri L Wagstaff. Machine Learning that Matters. In *Proceedings of the 29th International Conference on Machine Learning*, pages 529–536, Edinburgh, Scotland, 2012.

J. Williamson. A dynamic interaction between machine learning and the philosophy of science. pages 539–549, 2004. ISSN 0924-6495. doi: 10.1023/B:MIND.0000045990.57744.2b. URL http://kar.kent.ac.uk/7451/.

Jon Williamson. The philosophy of science and its relation to machine learning. *Scientific Data Mining and Knowledge Discovery*, pages 1–14, 2010. URL http://link.springer.com/chapter/10.1007/978-3-642-02788-8_4.

Lei Yu, Chris Ding, Steven Loscalzo, Lei Yu, Chris Ding, Chris Ding, Steven Loscalzo, and Steven Loscalzo. Stable feature selection via dense feature groups. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 08*, page 803, New York, New York, USA,

2008. ACM, ACM Press. ISBN 9781605581934. doi: 10.1145/1401890.1401986. URL http://portal.acm.org/citation.cfm?doid=1401890.1401986http://dl.acm.org/citation.cfm?doid=1401890.1401986.

Jiaming Zeng, Berk Ustun, and Cynthia Rudin. Interpretable Classification Models for Recidivism Prediction. *arXiv preprint arXiv:1503.07810*, (2014), 2015.

Jy Zou and Rp Adams. Priors for Diversity in Generative Latent Variable Models. *Nips*, pages 1–9, 2012. ISSN 10495258. URL https://papers.nips.cc/paper/4660-priors-for-diversity-in-generative-latent-variable-models.pdf.